Mon. Not. R. Astron. Soc. **000**, 1–19 (2011) Printed 4 March 2013 (MN LATEX style file v2.2)

BLOBCAT: Software to Catalogue Flood-Filled Blobs in Radio Images of Total Intensity and Linear Polarization

C. A. Hales[1,2⋆], T. Murphy[1,3,4], J. R. Curran[3], E. Middelberg[5], B. M. Gaensler[1,4], R. P. Norris[2,4]

[1] *Sydney Institute for Astronomy, School of Physics, The University of Sydney, NSW 2006, Australia*
[2] *CSIRO Astronomy & Space Science, PO Box 76, Epping, NSW 1710, Australia*
[3] *School of Information Technologies, The University of Sydney, NSW 2006, Australia*
[4] *ARC Centre of Excellence for All-sky Astrophysics (CAASTRO)*
[5] *Astronomisches Institut, Ruhr-Universität Bochum, Universitätsstr. 150, 44801 Bochum, Germany*

4 March 2013

ABSTRACT

We present BLOBCAT, new source extraction software that utilises the flood fill algorithm to detect and catalogue blobs, or islands of pixels representing sources, in two-dimensional astronomical images. The software is designed to process radio-wavelength images of both Stokes I intensity and linear polarization, the latter formed through the quadrature sum of Stokes Q and U intensities or as a byproduct of rotation measure synthesis. We discuss an objective, automated method by which estimates of position-dependent background root-mean-square noise may be obtained and incorporated into BLOBCAT's analysis. We derive and implement within BLOBCAT corrections for two systematic biases to enable the flood fill algorithm to accurately measure flux densities for Gaussian sources. We discuss the treatment of non-Gaussian sources in light of these corrections. We perform simulations to validate the flux density and positional measurement performance of BLOBCAT, and we benchmark the results against those of a standard Gaussian fitting task. We demonstrate that BLOBCAT exhibits accurate measurement performance in total intensity and, in particular, linear polarization. BLOBCAT is particularly suited to the analysis of large survey data.

Key words: methods: data analysis, statistical — techniques: image processing, polarimetric.

1 INTRODUCTION

In radio astronomy image analysis, for which approximations of Gaussian noise statistics and Gaussian source morphologies are suitable, much attention has been paid to least squares 2D elliptical Gaussian fitting routines (e.g. Condon 1997). Such routines, for example those implemented within the MIRIAD (Sault et al. 1995) and AIPS (Bridle & Greisen 1994) packages, are appropriate for source extraction when fitting parameters have been carefully inspected or constrained. However, when left unconstrained, the accuracy of these Gaussian fits may become degraded, requiring significant manual inspection overheads to identify poor fits and ensure high quality source extraction. Gaussian fitting routines may therefore be unsuited to the general analysis of large survey data.

In this work we seek to develop a robust alternative to Gaussian fitting by utilizing the flood fill algorithm (Lieberman 1978; Fishkin & Barsky 1985). In particular, we seek to develop a source extraction procedure that incorporates an accurate, objective, and automated method of background root-mean-square (rms) noise estimation, and to develop the first accurate method of source extraction for resolved sources in linear polarization. Additional factors motivating this work are described as follows.

First, a number of large radio surveys are planned for the near future, capitalising on upcoming new or substantially upgraded facilities such as ASKAP (Johnston et al. 2008; Deboer et al. 2009), MEERKAT (Jonas 2009), LOFAR (Rottgering et al. 2010), ALMA (Wootten & Thompson 2009; Hills et al. 2010), LWA (Ellingson et al. 2009), WSRT (Oosterloo et al. 2009), EVLA (Perley et al. 2011), and many others including VLBI networks and epoch of reionisation instruments. With these facilities will come a number of large surveys in both total intensity and linear polarization, for example EMU

⋆ E-mail: c.hales@physics.usyd.edu.au

(Norris et al. 2011), WODAN[1], MIGHTEE[2], POSSUM (Gaensler et al. 2010), and GALFACTS (Taylor & Salter 2010). The ability to catalogue objects within the large images produced by these surveys, with as little manual intervention as possible, will be key to maximising scientific output. We seek to develop a robust, automated method of source extraction that requires only the most complex of sources to be manually inspected.

Second, recent polarimetric studies have indicated an increase in the fractional polarization of faint extragalactic radio sources (e.g. Taylor et al. 2007; Subrahmanyan et al. 2010; Grant et al. 2010; Shi et al. 2010), which are difficult to reconcile with population modelling (O'Sullivan et al. 2008). We seek here to subject the process of polarization measurement to close scrutiny, and to provide the community with a measurement tool that has been assessed within a controlled testing environment.

And third, the flood fill algorithm underpins a number of existing source extraction routines, such as those available in the `CUPID`[3] (e.g. `CLUMPFIND` Williams et al. 1994) and `SExtractor` (Bertin & Arnouts 1996) packages. However, these routines are unable to measure flux densities without performing subsequent Gaussian (or similar) source fitting. Alternatively, the flood fill algorithm has been used without the subsequent least squares fitting step for the customised analysis of extended, non-Gaussian sources in total intensity (Murphy et al. 2007) and linear polarization (Heald et al. 2009). However, the raw flood fill algorithm as implemented in these works is not suitable for use with compact (unresolved or resolved Gaussian) sources, as their flux density measurements suffer from two significant systematic biases. In this work we describe how to correct for these biases in a robust manner, so as to enable the flood fill approach to handle both Gaussian and non-Gaussian sources.

We have implemented these bias corrections within a new flood fill program called `BLOBCAT`, which catalogues blobs in astronomical images. We use the term *blob* in an image-processing sense to represent an island of agglomerated pixels within a sea of noise, and to indicate that its properties are not inferred by fitting (e.g. least squares). We have designed `BLOBCAT` for use in radio astronomy, attempting to produce a program capable of encapsulating the entire measurement process between observational image and output catalogue.

This paper is organised as follows. In § 2 we describe the algorithms implemented within `BLOBCAT`, detailing required program inputs, including the minimal set required for operation, and output data products. In § 3 we assess `BLOBCAT`'s peak surface brightness (SB), integrated SB, and positional measurement performance. We investigate the program's ability to handle unresolved, resolved, and complex (non-Gaussian) sources in images of total intensity (Stokes I) and linear polarization (L or $L_{\rm RM}$; these terms are defined in § 2), and discuss issues regarding polarization bias. For comparison, we also assess the performance of a standard Gaussian fitting routine. In § 4 we discuss two examples of post-processing that may be required to make full use of `BLOBCAT`'s output catalogue; these are particularly relevant for data containing extended non-Gaussian, or multiple blended Gaussian, sources. In § 5 we present our summary and conclusions.

[1] http://www.astron.nl/radio-observatory/apertif-eoi-abstracts-and-contact-information

[2] Van der Heyden K., Jarvis M. J., 2010, MIGHTEE proposal to MEERKAT

[3] http://starlink.jach.hawaii.edu/starlink/CUPID

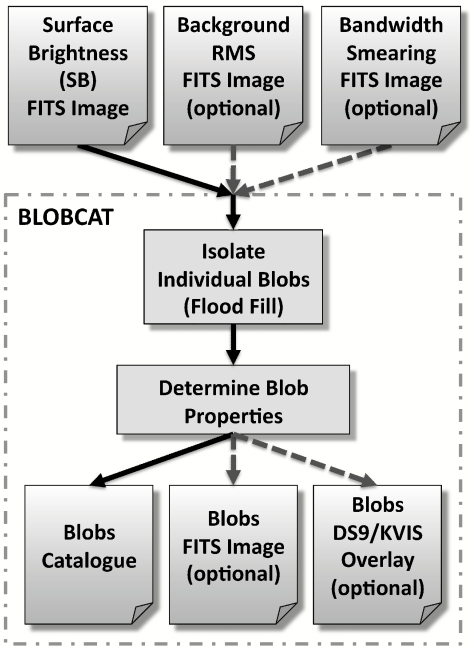

Figure 1. Overview of `BLOBCAT`.

2 HOW `BLOBCAT` WORKS

`BLOBCAT` is written in the scripting language `Python`. The program is designed to catalogue blobs in a two-dimensional (2D) input FITS (Pence et al. 2010) image of SB. To isolate blobs and determine their properties, `BLOBCAT` requires an estimate of the background rms noise and degree of bandwidth smearing at each spatial position (pixel) within the SB image. These two diagnostics may be provided to `BLOBCAT` as either uniform (spatially-invariant) values or, more generally, as 2D input FITS images that encode the more realistic scenario whereby noise and smearing properties vary with spatial position over the SB image.

An overview of `BLOBCAT` is presented in Fig. 1. In the following sections we describe the input images and their requirements (§ 2.1), the core flood fill algorithm used to isolate blobs (§ 2.2), the key morphological assumption (§ 2.3) and bias corrections (§ 2.4) applied to extract blob properties, the input arguments required to run `BLOBCAT` (§ 2.5), the output catalogue (§ 2.6), and the optional program outputs (§ 2.7).

2.1 Input Images

`BLOBCAT` requires up to three input FITS images, as outlined in Fig. 1. For flexibility, the images of background rms noise and bandwidth smearing are optional, and may instead be replaced by spatially-invariant input values.

2.1.1 Surface Brightness

`BLOBCAT` is designed to analyse blobs with positive SB. To detect negative blobs, the input SB image must be inverted before use. In this paper we focus on the analysis of blobs in images of total intensity and linear polarization (L or $L_{\rm RM}$). `BLOBCAT` may also be used to analyse images of Stokes Q, U, and V intensities, though we note that resolved sources exhibiting both positive and negative SB in these images will be incorrectly handled; we do not attempt to address the analysis of such sources here. We assume that blobs of interest in total intensity and linear polarization may be characterised by 2D elliptical Gaussians, though we do consider the treatment of non-Gaussian blobs later in § 4.2. Image pre-processing techniques to remove wide-spread extended features prior to the analysis of more compact sources may be required (e.g. Rudnick 2002; Rudnick & Brown 2009; Oppermann et al. 2011).

We assume that images of $L_{\rm RM}$ are produced following the application of rotation measure (RM) synthesis (Brentjens & de Bruyn 2005) and RMCLEAN (Heald et al. 2009) such that for each spatial pixel located at pixel coordinate (x, y), the polarized emission is obtained by taking the peak value in the cleaned Faraday dispersion function, namely

$$L_{\rm RM}(x,y) = \max(||F^{\rm cleaned}(x,y,\phi)||) \,, \tag{1}$$

where ϕ is Faraday depth. We note that this definition of $L_{\rm RM}$ assumes Faraday spectra along each pixel sightline consisting of no more than a single unresolved Faraday component (additional components will be ignored); analysis with more advanced models of $L_{\rm RM}$ are beyond the scope of this work. Analysis involving equation (1) is demonstrated, for example, by Heald et al. (2009) and Hales et al. (in preparation). Alternatively, images of standard linear polarization,

$$L(x,y) = \sqrt{Q(x,y)^2 + U(x,y)^2} \,, \tag{2}$$

may be used. See Leahy & Fernini (1989) and Vaillancourt (2006) for statistical properties of L, and Hales et al. (2012) for statistical properties of both L and $L_{\rm RM}$. For simplicity in subsequent discussion, we neglect the pixel coordinate notation (x, y) affixed to all spatially variable parameters, unless required for clarity.

2.1.2 Background RMS Noise

If position-dependent rather than spatially-invariant blob detection thresholds are required, then a background rms noise image must be specified. The user is required to independently construct a suitable noise map for their SB image, for example using the rms estimation algorithm implemented within the `SExtractor` package (Bertin & Arnouts 1996; Holwerda 2005).

Despite having been originally developed for the analysis of optical photographic plate and CCD data, `SExtractor` has been found to be reliable when generating noise maps from radio data (Bondi et al. 2003; Huynh et al. 2005). `SExtractor` determines the rms noise at each spatial pixel in an image by extracting the distribution of pixel values within a local mesh, iteratively clipping the most deviant values until convergence is reached at $\pm 3\sigma$ about the median. The choice of mesh size (in pixel^2) is very important. If it is too small, the local rms estimate may be biased due to lack of statistically independent measurements or overestimated due to the presence of real sources. If it is too large, any true small-scale variations in local rms noise may be washed out. At least $N_b = 80$ independent resolution elements (beams) per mesh area are required in order to reduce the uncertainty in estimates of local rms noise to below $\{[1 + 0.75/(N_b - 1)]^2[1 - N_b^{-1}] - 1\}^{0.5} = 8\%$ (using an approximation to the uncertainty of the standard error estimator, suitable for $N_b > 10$; p. 63, Johnson & Kotz 1970). The mesh area, $H_{\rm mesh}$, may be calculated according to

$$H_{\rm mesh} = \frac{N_b}{\bar{d}} \Omega_b \,, \tag{3}$$

where

$$\Omega_b = \frac{\pi}{4 \ln 2} \Theta_{\rm maj} \, \Theta_{\rm min} \tag{4}$$

is the beam volume for a 2D elliptical Gaussian with full-width at half-maximum (FWHM) along the major and minor axes given by $\Theta_{\rm maj}$ and $\Theta_{\rm min}$, respectively, and where $\bar{d} = \pi/\sqrt{12}$ is the densest lattice packing for congruent copies of any convex shape (e.g. circles, ellipses; Pach & Agarwal 1995). It is customary in physical sciences to treat rms noise[4] values, such as those reported by `SExtractor`, as standard errors in order to boost noise estimates in regions where extended non-signal features are present; namely by defining that $\sigma_z = (z_{\rm rms})_{\rm SExtractor}$. In other words, by using rms noise estimates to calculate local signal-to-noise ratio (SNR) thresholds for blob detection, it is possible to take into account not only local variations in image sensitivity, but also the possible presence of DC offsets due to artefacts (e.g. sidelobes). For this reason we recommend the method of using `SExtractor` or a similar package to estimate noise over the method of simply estimating it from, say, Stokes V because it can take into account features in the data that may be missed by more theoretically motivated expectations. The procedure described above, incorporating equation (3), may be easily automated to provide objective estimates of rms noise for any noise-dominated image.

Finally, we note that the `SExtractor` procedure above is suitable for determining the rms noise in images of Stokes I, Q, U, or V, but not $L_{\rm RM}$ (nor L). Instead, to determine $\sigma_{\rm RM}$ at each spatial location in $L_{\rm RM}$, `SExtractor` should be run on each constituent Q_i and U_i image in each i'th of T frequency channels to obtain $\sigma_{Q,i}$ and $\sigma_{U,i}$. These in turn may then be combined using weighted least squares as (Hales et al. 2012)

$$\sigma_{\rm RM} = \left[\xi \sum_{i=1}^{T} \frac{1}{0.2 \min\left(\sigma_{Q,i}^2, \sigma_{U,i}^2\right) + 0.8 \max\left(\sigma_{Q,i}^2, \sigma_{U,i}^2\right)} \right]^{-\frac{1}{2}} , \tag{5}$$

where the term ξ represents the correlation correction factor defined by equation (23) from Hales et al. (2012).

2.1.3 Bandwidth Smearing

If corrections for position-dependent bandwidth smearing (chromatic aberration) are required, then an image detailing the degree of smearing at any location within the SB image must be specified. Bandwidth smearing is due to

[4] The definition of rms noise is $z_{\rm rms}^2 = \bar{z}^2 + \sigma_z^2$.

the finite bandwidth of frequency channels, resulting in a radially-dependent convolution (smearing) that worsens as a function of positional offset from the phase tracking centre of a single-pointed radio observation (Condon et al. 1998; Bridle & Schwab 1999). The effect is to decrease the peak SB and to increase the observed size of sources without affecting their integrated SB. Bandwidth smearing needs to be carefully accounted for in mosaics consisting of multiple overlapped pointings. This is because any location in a mosaicked image, even one situated over a pointing centre, may include multiple contributions from adjacent pointings in which bandwidth smearing is significant (Ibar et al. 2009). The bandwidth smearing image input to `BLOBCAT` should map out the ratio between the observed smeared peak SB, S_p, and the true unsmeared peak SB, $S_p^{\rm BWS}$, for all spatial positions within the SB image (using notation consistent with that introduced later in this work). We denote the local degree of bandwidth smearing as

$$\varpi = \frac{S_p}{S_p^{\rm BWS}} \quad (\leqslant 1) \ . \tag{6}$$

2.1.4 General Requirements

All images input to `BLOBCAT` must have the same dimensions and be located on the same pixel grid; for cataloguing purposes we require that the primary image world coordinate system is expressed in equatorial coordinates (RA-Dec). In order to measure fitted Gaussian peaks to within 1%, at least 5 pixels per resolution element FWHM should be present (see Appendix A).

`BLOBCAT` does not calculate the Jacobian of the transformation between projection plane coordinates and native longitude and latitude (Calabretta & Greisen 2002). Instead, `BLOBCAT` requires that input images are gridded to an equal-area projection, so as to ensure that sky area per pixel is preserved. `BLOBCAT` supports both zenithal equal-area (`ZEA`) projection (the premier scheme for a hemisphere) and Hammer-Aitoff (`AIT`) equal-area projection (suitable for all-sky images) (Calabretta & Greisen 2002). Failure to use an equal-area projection will lead to systematic biases in `BLOBCAT`'s extracted flux densities and visibility area (sky density) calculations (see § 2.6). However, there are two common situations where this equal-area requirement may be relaxed. The first is when measuring flux densities for unresolved sources by obtaining their peak pixel or fitted peak value (cf. Appendix A). The second involves the use of images with non-equal-area projections; for example, the North-celestial-pole (`NCP`) projection (Greisen 1983). For such images, flux density measurements for resolved sources, which require integration over SB (i.e. over pixels), will only be suitable for sources situated close to the image reference point where distortion effects are minimal (Calabretta & Greisen 2002). To enable such analysis, `BLOBCAT` also supports images in `NCP` projection or the more general slant orthographic (`SIN`) projection. Regridding of input images to one of the `ZEA`, `AIT`, `NCP`, or `SIN` projection schemes may be computed using, for example, the `WCSLIB`[5] package. Finally, we remark that equal-area projections do not preserve shape; it is not possible to conserve both angles and areas when mapping portions of a sphere to a plane.

[5] http://www.atnf.csiro.au/people/mcalabre/WCS/wcslib/

2.2 Flood Fill Algorithm

`BLOBCAT` uses the flood fill, or thresholding, algorithm (Lieberman 1978; Fishkin & Barsky 1985; Sonka, Hlavac & Boyle 2008) to isolate individual blobs (islands) of pixels from within a SNR map. The SNR map is formed by taking the pixel-by-pixel ratio between the input SB and background rms noise images. In units of dimensionless SNR, we denote the threshold for detecting blobs as T_d and the cut-off threshold for flooding down to as T_f. By applying thresholds in the SNR map rather than the SB image, local variations in sensitivity can be accommodated. We do not take into account bandwidth smearing at this initial flooding stage (see § 2.6 below). We have implemented the highly optimised flood fill algorithm from Murphy et al. (2007) within `BLOBCAT`, which operates as follows.

(i) Locate all pixels in the SNR map that have value $\geqslant T_d$, including those pixels that would meet this detection threshold if it were not for pixellation attenuation (see Appendix A and comments below).

(ii) Form blobs about each of these pixels by 'flooding' adjacent pixels that have value $\geqslant T_f$.

(iii) For each isolated blob, perform bias corrections (§ 2.4) and catalogue properties (§ 2.6).

We denote the peak SB observed within the peak pixel for each blob by $S_p^{\rm OBS}$ (with units Jy beam^{-1}), and the resulting observed peak SNR by $A^{\rm OBS} = S_p^{\rm OBS}/\sigma$. To minimise the attenuating effect of pixellation on $S_p^{\rm OBS}$, `BLOBCAT` calculates a fitted peak SB for each blob by applying a 2D parabolic fit to a 3×3 pixel array about the raw peak, as described in Appendix A. We denote this fitted peak by $S_p^{\rm FIT}$, and the resulting fitted peak SNR by $A^{\rm FIT} = S_p^{\rm FIT}/\sigma$. We denote measurements of integrated SB by $S_{int}^{\rm OBS}$ (with units Jy), which are obtained for each blob by summing their flooded pixel intensities and dividing by the beam volume (Ω_b).

`BLOBCAT` attempts to perform its internal calculations, as described in the following sections, using the fitted peak quantities $S_p^{\rm FIT}$ and $A^{\rm FIT}$. However, if $S_p^{\rm FIT} < S_p^{\rm OBS}$, as may occur for heavily pixellated images (namely, for small values of N_α and N_δ as defined in Appendix A), then for consistency `BLOBCAT` sets $S_p^{\rm FIT} = S_p^{\rm OBS}$ (and thus $A^{\rm FIT} = S_p^{\rm OBS}/\sigma$) to ensure that blobs with $S_p^{\rm FIT} < T_d$ yet $S_p^{\rm OBS} > T_d$ are not unfairly rejected from the output catalogue. For notational simplicity in subsequent discussion we will use the superscript `OBS` to refer to both unfitted and fitted peak quantities; we will not distinguish between `OBS` and `FIT` quantities unless required for clarity.

We now turn to the key morphological assumption used to infer physical properties of these isolated blobs from their raw observed measurements.

2.3 Blob Morphology Assumption

In aperture synthesis imaging, individual resolution elements are described by the morphology of the dirty beam (the Fourier transform of the sampling distribution). Provided that the central core of the dirty beam can be suitably

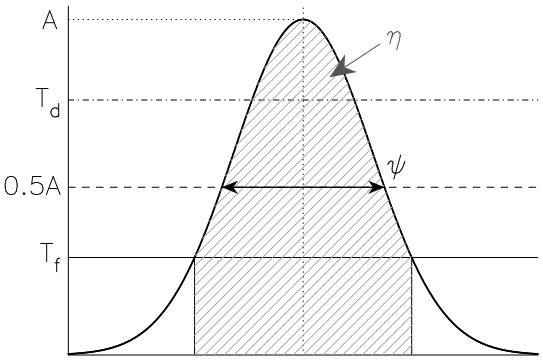

Figure 2. Flood fill algorithm applied to a noise-free 2D elliptical Gaussian blob with peak SNR A. The detection threshold is T_d. The blob is flooded from the peak down to the detection threshold T_f. Flood fill can only measure a fraction of the blob's total volume, η (equation (16)), as indicated by the shading. The width of the blob at $A/2$ (the FWHM) is ψ.

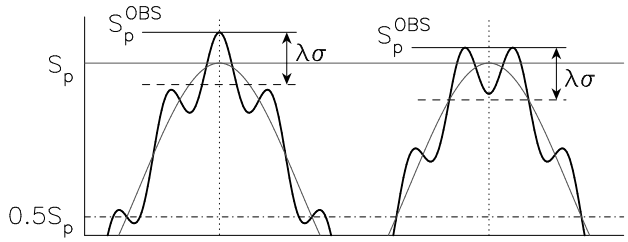

Figure 3. Idealised representation of the positive bias encountered when measuring the peak SB of a resolved Gaussian blob embedded in noise. Shown are two resolved Gaussian blobs, each with (true) peak SB S_p and 7 resolution elements per FWHM. For visual and conceptual simplicity, noise is represented by a sine wave and it is assumed that a large number of pixels populate each resolution element (such that pixellation effects may be ignored; i.e. $S_p^{\rm FIT} = S_p^{\rm OBS}$). Two equally likely noise superpositions are shown. The left blob encounters a positive noise contribution to its peak SB while the right blob encounters a negative noise contribution (trough). In both cases the observed peak SB overestimates the true peak SB, leading to a systematic positive bias for resolved sources. `BLOBCAT` corrects for this bias with equation (14), as parameterised by the area sliced at $\lambda\sigma$ below the observed peak. If λ is too small, the bias correction itself may become biased due to volatility in the small area sliced, as illustrated.

approximated by an elliptical Gaussian, then individual resolution elements in the resulting images can be described by 2D elliptical Gaussians. In other words, point sources will appear as Gaussians in an image.

In `BLOBCAT` we assume that each isolated blob is described by a 2D elliptical Gaussian characterised by a peak SNR, A, and representative major and minor FWHMs ψ_r and ψ_s, respectively (representative because these FWHMs are never individually measured, as we discuss shortly). In § 3.3 and § 4.2 we discuss situations where this assumption of Gaussian blob morphology is poor. The general equation for a 2D elliptical Gaussian, located at the origin of an arbitrary coordinate frame (r, s) that is aligned with the major/minor axes, is given by

$$f(r,s) = A\exp\left[-4\ln(2)\left(\frac{r^2}{\psi_r^2}+\frac{s^2}{\psi_s^2}\right)\right] . \qquad (7)$$

This equation is valid for Gaussian blobs in noise-free images of either total intensity or linear polarization. The volume of this 2D Gaussian is

$$\Omega_G = \frac{\pi A}{4\ln 2}\psi_r\psi_s . \qquad (8)$$

This general setup, including detection thresholds as defined in § 2.2, is shown in Fig. 2.

2.4 Blob Bias Corrections

`BLOBCAT` applies two important corrections to each isolated Gaussian blob in order to prevent systematic biases from affecting its peak and integrated SB measurements. These corrections account for:

(i) The positive peak surface brightness bias exhibited by $S_p^{\rm OBS}$ for resolved blobs; and
(ii) The negative integrated surface brightness bias exhibited by $S_{int}^{\rm OBS}$ caused by the limited blob volume accessible to flooding before the cut-off threshold T_f is reached.

2.4.1 Peak Surface Brightness Bias

An illustration outlining the need for the first correction is presented in Fig. 3. To understand this bias and how to correct for it, we first examine the following experiment. Consider for simplicity that blobs are represented by tophat functions rather than 2D elliptical Gaussians, that images are produced with one pixel per resolution element, and that noise is Gaussian in character. Noise is always resolved on the same spatial scale as unresolved sources. Therefore, the peak SB of an unresolved blob, here observed as the magnitude of a single pixel, will be affected by a single noise sample which may be positive or negative. For an ensemble of such unresolved blobs, each with identical true peak SB but different noise sample, the average observed peak SB will be an unbiased tracer of the true peak SB. Now consider a resolved tophat blob, over which M independent noise samples will be present. The observed peak SB of this resolved blob will depend on the maximum of M independent noise samples, rather than $M = 1$ for an unresolved blob. Thus the more resolved the blob becomes, the larger M becomes, and the less likely it is that a negative noise sample will be selected as the observed peak SB. The average observed peak SB for an ensemble of identically resolved blobs will therefore be positively biased from its true value. Before returning to 2D elliptical Gaussians, we will describe how to correct for this positive bias in the context of order statistics using the simpler tophat blob morphology.

For a sample of M independent and identically-distributed variates $X_1, X_2, \ldots, X_M$ ordered such that $X_{(1)} < X_{(2)} < \ldots < X_{(M)}$ (using notation X_j for unordered variates and $X_{(j)}$ for ordered variates), then $X_{(k)}$ is known as the k'th order statistic and $X_{(M)} = \max(X_j)$. If X has density function $f(X)$ and distribution function $F(X)$, then David & Nagaraja (2003) give the density function for $X_{(k)}$ as

$$f(X_{(k)}) = \frac{M!}{(k-1)!\,(M-k)!}\,f(X)\,[F(X)]^{k-1}\,[1-F(X)]^{M-k} . \qquad (9)$$

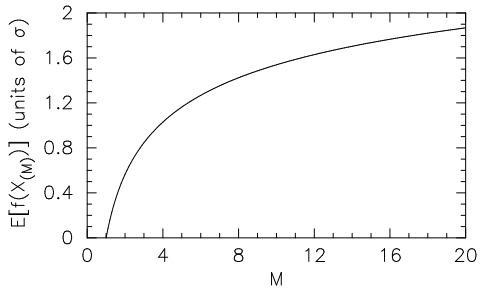

Figure 4. Expectation value in noise units of σ for the largest of M independent Gaussian variates (equation (12)). The expectation value is 0 for $M = 1$. A polynomial fit to the curve is given by equation (14).

The density function for the maximum of M independent Gaussian variates with variance σ^2 is obtained from equation (9) by setting $k = M$, giving

$$f(X_{(M)}) = \frac{M}{\sigma\sqrt{2\pi}} \exp\left(-\frac{X^2}{2\sigma^2}\right) \left\{\frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{X}{\sigma\sqrt{2}}\right)\right]\right\}^{M-1} , \quad (10)$$

where erf is the error function defined by

$$\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} dt . \quad (11)$$

The expectation value for equation (10) is given by

$$\mathrm{E}\left[f(X_{(M)})\right] = \int_{-\infty}^{\infty} f(X_{(M)})\ dX , \quad (12)$$

which is plotted for a range of M samples in Fig. 4. Equation (12) represents the average positive bias existing between measurements of observed peak SB and true peak SB for a tophat blob. Given measurement of M, namely the number of independent resolution elements present over the extent of the blob, an estimate for the bias can be obtained. The bias is most pronounced for low SNR resolved blobs; for a tophat blob of extent ~ 4 resolution elements, the bias for a 5σ blob is $\sim 1.0\sigma/5\sigma = 20\%$ (see Fig. 4).

We now return to the scenario whereby blobs are assumed to represent 2D elliptical Gaussians. Instead of obtaining M from the full spatial extent of a tophat blob, M needs to be estimated from the observable properties of a 2D Gaussian embedded in noise. In `BLOBCAT` we estimate M by approximating that the relevant number of independent resolution elements contributing to the positive bias can be extracted from the cross-sectional area contained within a slice of constant-SNR at a few σ below the peak, as parameterised by λ in Fig. 3. `BLOBCAT` measures the cross-sectional area for each blob at SNR $= (A^{\mathrm{OBS}} - \lambda)$, which we denote H_λ, by flooding from the peak to this threshold and simply counting the number of pixels present. M is then estimated using (cf. equation (3))

$$M = \frac{\bar{d}}{\Omega_b} H_\lambda . \quad (13)$$

To determine the positive bias between S_p^{OBS} and S_p for resolved blobs, `BLOBCAT` uses the following fifth-order polynomial fit to the curve in Fig. 4 to form a simple lookup table (rather than solving for equation (12)),

$$M = 1 + \sum_{i=1}^{5} a_i \beta^i , \quad (14)$$

where

$$\beta = \mathbf{E}\left[f(X_{(M)})\right] \approx \frac{S_p^{\mathrm{OBS}}}{S_p} \left(= \frac{A^{\mathrm{OBS}}}{A}\right) , \quad (15)$$

and where $a_1 = 0.89$, $a_2 = 0.27$, $a_3 = 3.75$, $a_4 = -3.67$, and $a_5 = 1.61$.

To illustrate the constraints on selecting λ, imagine trying to correct the raw observed peak SB for a resolved Gaussian blob, detected with peak SNR = 50, by arbitrarily defining that the relevant spatial extent be measured at $\lambda = 20$. Choosing M in this way will overestimate the peak's positive bias, because not even a 10σ noise spike located at the SNR = 30 contour of the blob could be mistaken for the true peak. Alternatively, choosing too small a value of λ will not only underestimate the peak bias in the opposite manner to above, but also render M vulnerable to additional negative bias due to H_λ being fooled (limited in spatial extent) by noise troughs near the blob's peak.

We performed simulations to empirically determine the most suitable range of values for λ. We found that choosing $\lambda = 3.5$ best corrected for the positive bias exhibited by S_p^{OBS} for resolved blobs in images of either total intensity or linear polarization (L or L_{RM}). We discuss the simulations used to determine this optimum λ, as well as the general performance of the peak SB bias correction from equation (14), in § 3.

2.4.2 Integrated Surface Brightness Bias

To prevent the flood fill algorithm from cascading into noise features adjacent to real blobs, flooding is terminated at the cut-off threshold, T_f. The integrated SB measured for each blob, S_{int}^{OBS}, therefore underestimates the true integrated SB, S_{int}, because only a limited fraction of the total volume for each blob is ever directly accessed. We denote this fraction η, as indicated in Fig. 2.

By integrating the volume flooded between A (true peak SNR) and T_f for a 2D elliptical Gaussian blob, and dividing this result by the total volume of the blob (equation (8)), the fraction of flooded volume η is found to be

$$\eta = \left(\mathrm{erf}\sqrt{-\ln\frac{T_f}{A}}\right)^2 . \quad (16)$$

`BLOBCAT` corrects the observed integrated SB for each detected blob (regardless of blob dimension) by simply dividing by η, namely

$$S_{int} = \frac{S_{int}^{\mathrm{OBS}}}{\eta} . \quad (17)$$

It is important to note that A in equation (16) is the true peak SNR. For resolved blobs, the peak bias correction from equation (14) needs to be applied first, so as to debias the observed peak SNR, A^{OBS}, and return an estimate for the unbiased peak SNR, A. The effect of using uncorrected peak SNRs for resolved sources in equation (16) is demonstrated in § 3.

The choice of T_f affects the maximum volume that can be flooded within a faint blob. So as to recommend a minimum value, we performed simulations of integrated SB recovery for 2D elliptical Gaussian blobs embedded within images of total intensity and linear polarization; the details of these simulations are discussed in § 3. We incrementally reduced T_f in these simulations, seeking a balance between the measurement of as much volume as possible within faint blobs, and the need to avoid bias from potential over-flooding of neighbouring noise features.

In total intensity images for blobs as faint as $A = 5$, we found that a cut-off threshold of $T_f = 2.6$ was required in order to robustly flood as many true blob pixels as possible whilst avoiding over-flooding of adjacent non-blob (noise) pixels. In linear polarization images (L or $L_{\rm RM}$), non-Gaussian noise statistics typically limit detection thresholds to $T_d \gtrsim 6$ (Vaillancourt 2006; Hales et al. 2012). These images thus require higher flooding thresholds than those for total intensity; we note that a comparison between the average cross-sectional profile of a Gaussian blob embedded in images exhibiting Gaussian, L, and $L_{\rm RM}$ statistics is presented by Hales et al. (2012). In images of $L_{\rm RM}$ for blobs as faint as $A = 6$, we found that a cut-off threshold of $T_f = 4.0$ was suitable. We note that this value of T_f is dependent on the observational setup used to produce $L_{\rm RM}$. To determine an equivalent value of T_f for any L or $L_{\rm RM}$ image, a cut-off with equal statistical significance to our suggested $T_f = 4.0$ value should be calculated (e.g. see Hales et al. 2012).

For a detection threshold of $T_d = 5$ in an image of total intensity, equation (16) with $T_f = 2.6$ implies that the maximum correction factor for any blob is $1/\eta \lesssim 1.8$. In linear polarization, for a detection threshold of $T_d \sim 6$ and $T_f = 4.0$, the maximum correction factor is $1/\eta \lesssim 2.5$.

2.5 Program Inputs

If accurate error estimates are not immediately required, `BLOBCAT` does not requires many inputs to run. Preliminary analysis can be performed on a single input SB image by specifying three parameters: a background rms noise value (simply so that SNRs can be computed at any spatial location within the image), a blob detection SNR threshold (T_d), and a cutoff SNR threshold for flooding (T_f). However, to make full use of the output catalogue, particularly errors, additional input parameters are required. For completeness, we list all `BLOBCAT` input arguments in Appendix B.

2.6 Output Catalogue

`BLOBCAT` produces an output catalogue containing 41 entries for each detected blob. In this section we list and define these entries, which include final measurements of peak and integrated SB, corrected for bandwidth smearing and clean bias, errors, and the 'visibility' area for each blob. The catalogue entries, some of which require various `BLOBCAT` input arguments to be specified (see Appendix B), are as follows.

Column 1: `ID`
Blob identification number, ordered by decreasing observed peak SNR (see Column 26).

Column 2: `npix`
Number of flooded pixels comprising blob.

Columns 3-4: `x_p, y_p`
RA and Dec of peak pixel in pixel coordinates.

Columns 5-6: `RA_p, Dec_p`
RA and Dec of peak pixel in degrees.

Column 7: `RA_p_err`
Total position error in RA of peak pixel, which we define as

$$\sigma_\alpha = \sqrt{\sigma^2_{\alpha,\rm cal} + \sigma^2_{\alpha,\rm frame} + \sigma^2_{\alpha,\rm blob}} \, . \tag{18}$$

The first term, $\sigma^2_{\alpha,\rm cal}$, represents the positional uncertainty of the phase calibrator, for example with reference to the International Celestial Reference Frame, that was used to produce the SB image. The second term, σ^2_{frame}, represents the positional uncertainty of the image frame about the (assumed) position of the phase calibrator. Given that image positional errors correspond to Fourier-plane phase errors, $\sigma^2_{\rm frame}$ may be estimated by measuring $\sigma_{\rm SEM}$, the standard error of the mean (SEM) of the variation in the phase corrections resulting from phase self-calibration[6] (Cornwell & Fomalont 1999). By estimating the fraction of a resolution element by which positions may be in error as $\sigma_{\rm SEM}/180^\circ$, `BLOBCAT` estimates the frame error as

$$\sigma_{\alpha,\rm frame} \approx \frac{1}{\sqrt{2}} \frac{\sigma_{\rm SEM}}{180^\circ} \Theta_\alpha \, , \tag{19}$$

where the factor of $\sqrt{2}$ projects the 2D SEM along one of two orthogonal axes, and where Θ_α is the projected resolution along the RA-axis. Θ_α is given by

$$\Theta_\alpha = \frac{\Theta_{maj}\,\Theta_{min}}{\sqrt{\left(\Theta_{maj}\cos\chi\right)^2 + \left(\Theta_{min}\sin\chi\right)^2}} \, , \tag{20}$$

where χ is the position angle of the major axis East of North. The third term, $\sigma^2_{\alpha,\rm blob}$, encapsulates positional error due to the significance of the blob detection, which we define for reasons described later in § 3.1.3 and § 3.2.3 as

$$\sigma_{\alpha,\rm blob} \approx \frac{1}{1.4\,A} \Theta_\alpha \, . \tag{21}$$

Column 8: `Dec_p_err`
Total position error in Dec of peak pixel, which we define in a similar manner to equation (18) as

$$\sigma_\delta = \sqrt{\sigma^2_{\delta,\rm cal} + \sigma^2_{\delta,\rm frame} + \sigma^2_{\delta,\rm blob}} \, , \tag{22}$$

where

$$\sigma_{\delta,\rm frame} \approx \frac{1}{\sqrt{2}} \frac{\sigma_{\rm SEM}}{180^\circ} \Theta_\delta \, , \tag{23}$$

$$\sigma_{\delta,\rm blob} \approx \frac{1}{1.4\,A} \Theta_\delta \, , \tag{24}$$

and where the projected resolution along the Dec-axis is given by

$$\Theta_\delta = \frac{\Theta_{maj}\,\Theta_{min}}{\sqrt{\left(\Theta_{maj}\sin\chi\right)^2 + \left(\Theta_{min}\cos\chi\right)^2}} \, . \tag{25}$$

[6] Note that regardless of whether or not self-calibration phase corrections are applied to the visibility (Fourier) data prior to final imaging (i.e. it is possible to calculate the required phase corrections without applying them), the systematic positional offset between the image frame and the phase calibrator can be characterised by the SEM of the phase corrections (e.g. Hales et al. 2009).

Columns 9-10: `x_c, y_c`
RA and Dec of area (unweighted) centroid in pixel coordinates,

$$(x_c, y_c) = \frac{\sum_{i=1}^{\texttt{npix}} \mathbf{x}_i}{\texttt{npix}} \,, \tag{26}$$

where $\mathbf{x}_i = (x_i, y_i) \in$ blob.
Columns 11-12: `RA_c, Dec_c`
RA and Dec of unweighted centroid in degrees.
Column 13: `cFlag`
Centroid flag. If (x_c, y_c) is located within a flooded pixel, then cFlag = 1; otherwise cFlag = 0.
Columns 14-15: `x_wc, y_wc`
RA and Dec of SNR-weighted centroid in pixel coordinates,

$$(x_{wc}, y_{wc}) = \frac{\sum_{i=1}^{\texttt{npix}} \mathbf{x}_i \, A^{\rm OBS}(\mathbf{x}_i)}{\sum_{i=1}^{\texttt{npix}} A^{\rm OBS}(\mathbf{x}_i)} \,. \tag{27}$$

Columns 16-17: `RA_wc, Dec_wc`
RA and Dec of SNR-weighted centroid in degrees.
Column 18: `wcFlag`
Weighted centroid flag. If (x_{wc}, y_{wc}) is located within a flooded pixel, then wcFlag = 1; otherwise wcFlag = 0. If wcFlag = 1, then `RA_wc` and `Dec_wc` from Columns 16-17 above are the formal position of the blob. If wcFlag = 0, the blob is likely to be significantly non-Gaussian; the weighted-centroid position may not be suitable for formal cataloguing purposes. Manual inspection, or formal cataloguing using the raw peak pixel or area centroid positions, may be required.
Columns 19-22: `x_min, x_max, y_min, y_max`
Minimum and maximum pixel coordinate in RA (x) and Dec (y) within blob.
Column 23: `rms`
Rms noise, σ, at position of peak pixel.
Column 24: `BWScorr`
Bandwidth smearing correction, $1/\varpi$ (from equation (6)).
Column 25: `M`
Number of independent resolution elements from equation (13). M is used in equation (14) to correct for the positive peak bias exhibited by resolved blobs. To prevent this bias correction from being applied to noise-affected unresolved blobs (i.e. where the number of pixels flooded is artificially boosted due to a connected noise feature), `BLOBCAT` only applies the correction to those blobs with $M \geqslant 1.1$; the suitability of this value was determined empirically.
Column 26: `SNR_OBS`
Observed (raw) SNR, $A^{\rm OBS} = S_p^{\rm OBS}/\sigma$.
Column 27: `SNR_FIT`
Fitted SNR, $A^{\rm FIT} = S_p^{\rm FIT}/\sigma$.
Column 28: `SNR`
SNR, A, corrected for peak bias (equation (14)).
Column 29: `S_p_OBS`
Observed (raw) peak SB, $S_p^{\rm OBS}$.
Column 30: `S_p_FIT`
Fitted peak SB, $S_p^{\rm FIT}$, obtained using a 2D parabolic fit to a 3×3 pixel array about the raw peak pixel (x_p, y_p). If $S_p^{\rm FIT} < S_p^{\rm OBS}$, then `BLOBCAT` sets $S_p^{\rm FIT} = S_p^{\rm OBS}$ so as to use the more accurate measurement (see Appendix A and § 2.2).
Column 31: `S_p`
Peak SB, S_p, corrected for peak bias (equation (14)).
Column 32: `S_p_CB`
Peak SB corrected for peak bias and clean bias, $S_p^{\rm CB}$. Clean bias is a deconvolution effect that redistributes SB from real blobs to noise peaks, systematically reducing the observed SB of blobs independent of their SNR (Condon et al. 1998). The effect is more pronounced for observations with poor Fourier-plane coverage. Given the degree of clean bias present in the SB image, $\Delta S^{\rm CB}$ ($\geqslant 0$ Jy beam^{-1}), `BLOBCAT` makes the following correction,

$$S_p^{\rm CB} = S_p + \Delta S^{\rm CB} \,. \tag{28}$$

Column 33: `S_p_CBBWS`
Peak SB corrected for peak bias, clean bias and bandwidth smearing, $S_p^{\rm CB,BWS}$. Using the input value of ϖ (equation (6)), `BLOBCAT` corrects for bandwidth smearing with

$$S_p^{\rm CB,BWS} = \frac{S_p^{\rm CB}}{\varpi} \,. \tag{29}$$

This is the final reported value of the blob's peak SB, to be used for post-processing.
Column 34: `S_p_CBBWS_err`
Error in corrected peak SB, which we define as

$$\sigma_{S_p^{\rm CB,BWS}} = \Big[\left(\Delta S^{\rm ABS} S_p^{\rm CB,BWS}\right)^2 + \left(\Delta S^{\rm PIX} S_p^{\rm CB,BWS}\right)^2 + \left(\frac{\sigma}{\varpi}\right)^2 \Big]^{\frac{1}{2}} \,, \tag{30}$$

where $\Delta S^{\rm ABS}$ is the absolute calibration error of the SB image and $\Delta S^{\rm PIX}$ is the pixellation uncertainty (see Appendices A and B). The suitability of this error in linear polarization is discussed in § 3.2.
Column 35: `S_int_OBS`
Observed (raw) integrated SB, $S_{int}^{\rm OBS}$.
Column 36: `S_int_OBSCB`
Observed integrated SB corrected for clean bias, given by

$$S_{int}^{\rm OBS,CB} = S_{int}^{\rm OBS} + \frac{\texttt{npix}\ \Delta S^{\rm CB}}{\Omega_b} \,. \tag{31}$$

This value may be useful for non-Gaussian blobs (see § 3.3).
Column 37: `S_int`
Integrated SB, S_{int}, calculated by application of blob volume correction (equation (17)) to $S_{int}^{\rm OBS}$.
Column 38: `S_int_CB`
Integrated SB corrected for clean bias, $S_{int}^{\rm CB}$, calculated by application of blob volume correction (equation (17)) to $S_{int}^{\rm OBS,CB}$. This is the final reported value of the blob's integrated SB, to be used for post-processing (though see comments in § 3.3).
Column 39: `S_int_CB_err`
Error in corrected integrated SB, which we define in a similar manner to `S_p_CBBWS_err` (see also § 3.1) as

$$\sigma_{S_{int}^{\rm CB}} = \sqrt{(\Delta S^{\rm ABS} S_{int}^{\rm CB})^2 + \sigma^2} \,. \tag{32}$$

The suitability of this error in linear polarization is discussed in § 3.2.
Column 40: `R_EST`
Size estimate of detected blob, $R^{\rm EST}$, in units of the sky area covered by an unresolved Gaussian blob with the same peak SB, taking into account local bandwidth smearing. To derive this estimate we first focus on an unresolved Gaussian blob with FWHM Θ, as defined by the image resolution, and peak SB S_p, as measured from the detected blob. For this unresolved blob, the relationship between its full width at

a fraction T_f/A of its peak SB, which we denote φ, and its FWHM is given by

$$\varphi = \Theta\sqrt{\log_2 \frac{A}{T_f}} \; . \tag{33}$$

To calculate $R^{\rm EST}$ we take the ratio between the measured area of the detected blob, $H_{\rm blob}$, and the area of an ellipse with axes defined by equation (33). When the broadening effect of bandwidth smearing is included into this ratio, we get

$$R^{\rm EST} = H_{\rm blob}\left(\frac{\pi}{4}\frac{\Theta_{\rm maj}\Theta_{\rm min}}{\varpi}\log_2\frac{A}{T_f}\right)^{-1} \; . \tag{34}$$

The parameter $R^{\rm EST}$ is not intended to be used for quantitative analysis. In § 4 we discuss how $R^{\rm EST}$ may be used to flag blobs that exhibit potentially complex (non-Gaussian) morphologies for follow-up.

Column 41: `VisArea`

`BLOBCAT` can optionally calculate the fraction of visible sky area, namely the fraction of non-blank pixels assuming use of an equal-area projection, over which a blob detected at position (r, s) could have been detected within the SB image. This is known as the blob's visibility area. This area may be used, for example, to calculate a completeness correction when compiling source counts (e.g. Hales et al., in preparation). To calculate the visibility area, `BLOBCAT` takes into account spatial variations in both image sensitivity and bandwidth smearing. For non-blank pixels (x, y), the fraction of suitable sky area for detecting a blob with equal peak SB to that of a blob located at (r, s), where $r \in x$, $s \in y$, is obtained by counting the number of pixels that satisfy

$$\frac{T_d\,\sigma(x,y)}{\varpi(x,y)} \leqslant \frac{S_p(r,s)}{\varpi(r,s)} \; . \tag{35}$$

2.7 Optional Outputs

To aid visual inspection and post-processing of blobs, `BLOBCAT` can optionally produce two additional forms of output. The first is a modified SB FITS image in which all flooded pixels have been highlighted (reset to a large value; this value may be user-specified, see Appendix B). The second is an image overlay in `ds9` (Joye & Mandel 2003) or `Karma` (Gooch 1996) formats, for use with their respective `ds9` or `kvis` FITS viewers. The overlays present the identification number and boundaries in RA and Dec for each blob. To illustrate these two optional forms of output, an example output FITS image superposed with a `kvis` overlay is presented in Fig. 5. `BLOBCAT` may be easily modified to produce overlays in other suitable formats, for example through use of the `pywcs` wrapper to `WCSLIB`.

3 PERFORMANCE

We have carried out Monte Carlo simulations to investigate the performance of `BLOBCAT` in total intensity and linear polarization, as described in the following sections.

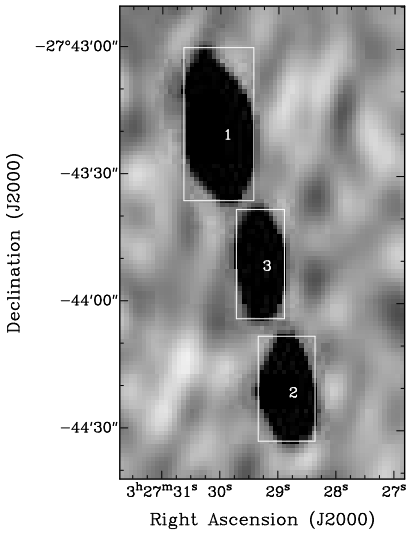

Figure 5. Output FITS image and `kvis` overlay as produced by `BLOBCAT`, illustrating how three blobs in the image are highlighted and identified (sample data from Norris et al. 2006).

3.1 Total Intensity

3.1.1 Simulation Setup

We tested `BLOBCAT` in total intensity by injecting Gaussian sources with peak SNRs between $5 - 100\sigma$ into images of Gaussian noise, inspecting the accuracy of the recovered SB and positional measurements. To compare `BLOBCAT`'s flood fill approach with that of standard Gaussian fitting, we also carried out these simulations using `IMFIT`, a widely used Gaussian source fitter from the `MIRIAD` package (Sault et al. 1995). Gaussian fitting routines such as `IMFIT` have been used by many surveys such as NVSS (Condon et al. 1998), Phoenix (Hopkins et al. 2003), and SUMMS (Mauch et al. 2003).

Two classes of source were tested, with the aim of demonstrating the virtues and limitations of `BLOBCAT`'s modified flood fill approach. The first were unresolved (point) sources, selected to demonstrate that flood fill algorithms need not be limited to the parameter space occupied by complex non-Gaussian sources. The second were highly (and somewhat pathologically) resolved Gaussian sources with FWHMs 5 times larger than the image resolution, probing parameter space where parameterised Gaussian fitting methods are optimal. We did not quantitatively address performance relating to non-Gaussian sources because of the lack of an obvious standardised test source; qualitative discussion of non-Gaussian blobs is presented in § 3.3.

We generated 125 independent samples per SNR bin using noise images produced as follows. To realistically characterise the noise environment present in images of total intensity, we obtained Stokes V data from an individual pointing of the mosaicked 1.4 GHz aperture synthesis observations of Norris et al. (2006). We imaged this Stokes V data using $1''$ pixels, and convolved to a final circular resolution with (FWHM) $\Theta = 14''$. We found this image to be free of sources and artefacts. Using `SExtractor` (see § 2.1.2), we modified this Stokes V image for use as a master noise image by enforcing zero mean and unit variance throughout

sub-regions of size 150 independent resolution elements. The noise image for each sample was then produced by extracting a randomly positioned thumbnail image from the master noise image, from a pool of over 150,000 choices.

For each sample we measured the injected source's peak SB, integrated SB, and position using both BLOBCAT and IMFIT. We executed IMFIT using unconstrained Gaussian fit parameters, imitating a blind survey. For input point sources, we also executed IMFIT using a constrained fit, fixing the source size to the image resolution. We then compared the output values for these different methods with their true input values. To prevent source misidentification, we checked that each recovered source extended over its true input location. We describe the results of these Monte Carlo simulations for SB measurements in § 3.1.2 and for positions in § 3.1.3.

3.1.2 Results and Discussion: Surface Brightness Measurements

We performed our total intensity Monte Carlo simulations for a range of flooding thresholds (T_f) and peak bias correction factors (λ), setting the detection threshold (T_d) as small as possible so as to limit the induction of sampling bias in the lowest SNR bins. For reasons outlined in § 2.4.1−2.4.2, we found that optimal SB recovery was obtained using $T_f = 2.6$ and $\lambda = 3.5$.

In Fig. 6 we present the SB results of our simulations, where we have executed BLOBCAT with the optimal T_f and λ values from above, we have executed IMFIT with unconstrained Gaussian fit parameters, and where we have used median statistics (Tukey 1977) to robustly prevent noise outliers from biasing intrinsic source extractor properties. The results obtained from executing IMFIT with constrained point source fits, using the same simulation data as for the unconstrained fits, are presented in Fig. 7. To put BLOBCAT's performance in perspective, we first discuss the results from IMFIT, starting with the unconstrained fits from Fig. 6.

The strength of IMFIT is its ability to perform least squares fitting in order to separate smooth underlying 2D elliptical Gaussians from superposed noise fluctuations. A key requirement of this process is that there are sufficient degrees of freedom (DOFs) to fit the position, peak SB, major and minor axis, and position angle parameters. Given that the number of DOFs is related to the number of independent resolution elements within the fitting region, it is to be expected that IMFIT will struggle to constrain multiple fit parameters for point-like input sources. This is reflected in the IMFIT results from Fig. 6, where the systematic bias in integrated SB measurements for point sources (top-right panel; $\gtrsim$15% at 5σ) demonstrates IMFIT's inability to simultaneously constrain peak SB and angular dimension parameters. For these point sources, which by definition have the dimensions of a single resolution element and therefore contain essentially one piece of information, namely their brightness, least squares fitting is easily coerced into including adjacent noise peaks into the fit. However, for resolved sources, which by definition extend over multiple independent resolution elements, least squares fitting becomes less likely to misinterpret noise features as true signal, and so becomes more accurate.

The systematic positive bias exhibited by IMFIT in its measurements of integrated SB for point sources leads to two systematic effects. First, given that the integrated to peak SB ratio is typically used to select which measure best characterises the flux density of a source (e.g. Huynh et al. 2005), the flux densities of faint sources will be systematically overestimated. Second, this ratio is often used to estimate deconvolved angular source sizes (e.g. Huynh et al. 2005), which too will become positively biased for faint sources. We comment on this ratio further in § 4.1.

We now turn to IMFIT's performance from Fig. 7. When

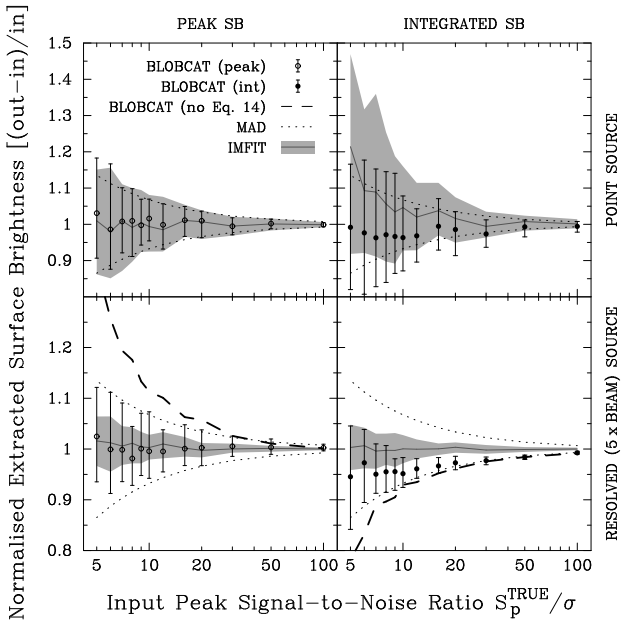

Figure 6. Performance of BLOBCAT (points) and IMFIT (shading) in total intensity for input unresolved (top row) and resolved (FWHM = 5 × image resolution; bottom row) Gaussian sources. Measurements of peak (left column) and integrated (right column) SB over a range of input peak SNRs are summarised by their median (points/curves), and first and third quartiles (whiskers/shading). Dashed curves trace median measurements resulting from exclusion of the peak bias correction for resolved sources (equation (14)). Fit parameters for IMFIT are unconstrained. For reference, expected random errors are indicated by the median absolute deviation (MAD $\approx 0.6745\sigma$; dotted curves). Note that the y-axis range differs between rows.

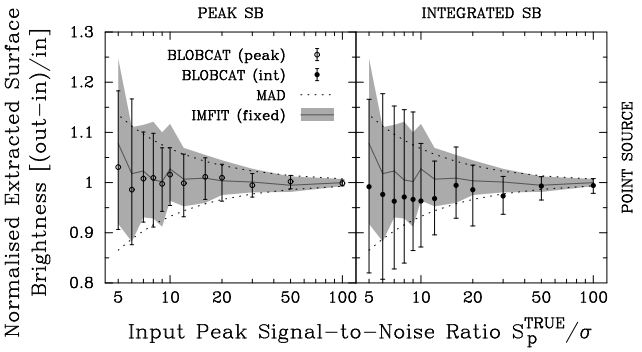

Figure 7. Reproduction of top row of Fig. 6, but here displaying IMFIT results for point source fits with angular dimensions fixed to the image resolution.

there is prior knowledge that a source is unresolved, IMFIT can be constrained to fit a point source, fixing its fitted dimensions to those of the image resolution. Comparing the results from Fig. 6 with those of Fig. 7, we find that the point source assumption reduces IMFIT's integrated SB bias, but does not completely eliminate it. Left behind is a marginal positive bias at low input SNR, caused by IMFIT's residual-minimisation strategy to pull fitted sources towards noise peaks that are directly adjacent to true source peaks. We comment further on measured positions in § 3.1.3.

Returning to the BLOBCAT results from Fig. 6, we find that the recovered peak and integrated SB measurements for point sources are systematically unbiased. This performance enhancement over IMFIT is due to the reduced influence that nearby noise features can exert over BLOBCAT's integrated SB measurements. Only directly connected noise features can affect flood fill, when the algorithm spills into adjacent noise peaks and is eventually limited by T_f, whereas strong noise peaks separated by a noise trough from the true source may be least square minimised by IMFIT as statistical fluctuations superposed on a resolved source.

For the resolved source investigated, IMFIT clearly outperforms BLOBCAT in avoiding integrated SB systematics. However, BLOBCAT's systematic underestimate is no worse than $\sim 5\%$, even for sources with peak SNR $= 5$. As indicated in Fig. 6, this underestimate would be more severe if the peak bias correction from equation (14) were neglected; failure to debias the peak SB causes equation (17) to underestimate the integrated SB. We attribute BLOBCAT's difficulty in collecting the full integrated SB for resolved sources to an analogous 'negative' version of our peak SB correction. As sources become more resolved, it becomes more likely that negative noise features may limit the spatial extent available for the flood fill algorithm to explore. This behaviour is not completely offset by positive noise features contributing to the spatial extent of sources, and so a bias is produced. Given how mild the resulting bias is, even for the pathologically resolved source tested, we do not attempt to correct for it within BLOBCAT.

To estimate the uncertainty in BLOBCAT's measurements of peak and integrated SB, we use equations (30) and (32). These errors are indicated by dotted lines in Fig. 6; we neglect the absolute calibration error (ΔS^{ABS}), and set the pixellation error (ΔS^{PIX}) to 0.5% (cf. Appendix A). We do not reduce the factor of σ in equation (32) by, for example, the square root of the number of independent resolution elements within the spatial extent of the source, as might be appropriate for methods that produce systematically unbiased integrated SB measurements. Instead, we define equation (32) in a similar manner to equation (30), so as to artificially account for BLOBCAT's systematic underestimate of integrated SB for resolved sources. In this way, the error estimates produced by BLOBCAT realistically encapsulate its true performance. Note that in practice, resolved sources will almost always be less resolved than for our simulated resolved source here. This implies that our catalogue error estimates are unlikely to underestimate true SB measurement errors.

3.1.3 Results and Discussion: Position Measurements

BLOBCAT catalogues three positions for each detected blob: the raw peak pixel, an area centroid using equation (26),

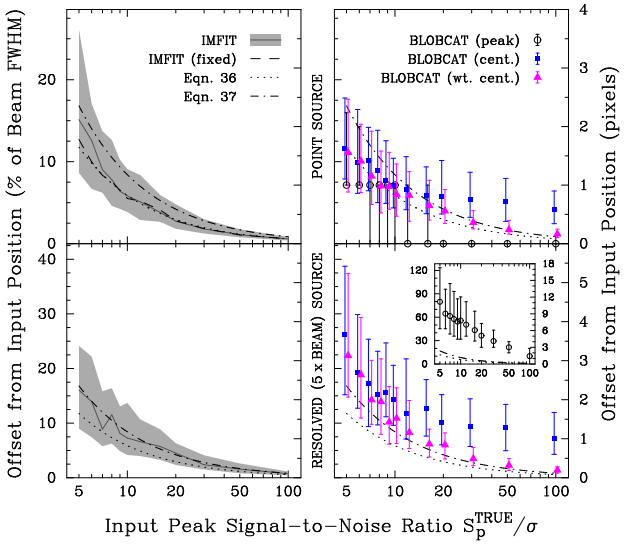

Figure 8. Accuracy of positions measured by IMFIT (shading; left column) and BLOBCAT (points for the peak pixel, centroid, and SNR-weighted centroid; right column) in total intensity for input unresolved (top row) and resolved (bottom row) Gaussian sources; median statistics are displayed (similar formalism to Fig. 6). The dashed curve (top-left panel) traces median measurements for constrained IMFIT point source fits with angular dimensions fixed to the image resolution. For reference, the dotted and dot-dashed curves (identical in each panel) indicate expected median positional offsets using equations (36) and (37), respectively. The left y-axis for each panel denotes position offset from the true input source position in units of the circular resolution FWHM ($\Theta = 14''$); the right y-axis denotes this offset in units of pixel width ($1''$). Note that the y-axis range differs between rows. For clarity, the bottom-right panel shows only centroid and SNR-weighted centroid measurements; the inset provides peak pixel measurements in a zoomed-out view of this panel.

and a SNR-weighted centroid using equation (27). In Fig. 8 we compare the accuracy of these measurements, as well position measurements from IMFIT, in recovering the true input positions for our simulated unresolved and resolved sources.

Fig. 8 indicates that of BLOBCAT's three position measurements, the weighted centroid is optimal for both unresolved and resolved Gaussian sources. The superior performance of the peak pixel position for unresolved sources is an artefact of injecting sources centred on a pixel; in general the performance of this position measure will be poorer. For resolved sources, the raw peak position is easily corrupted by the peak bias effect described earlier in § 2.4.1. For both unresolved and resolved Gaussian sources, the area centroid exhibits limited accuracy due to its lack of pixel weighting.

For faint unresolved sources, BLOBCAT's positions are more accurate than those of IMFIT's unconstrained Gaussian fits; IMFIT is limited in its accuracy due to its optimisation attempts to accommodate adjacent noise features through least squares minimisation. For the pathologically resolved source simulated, IMFIT's position measurements are more accurate than BLOBCAT's.

To estimate the uncertainty in BLOBCAT's weighted centroid positions, we first look to an uncertainty estimate for IMFIT. For plotting purposes, the median positional offset

exhibited by IMFIT can be estimated as the median of the quadrature sum of two zero-mean signals representing RA and Dec measurements with error σ_α (equation (18)) and σ_δ (equation (22)), respectively. By using a factor of $\sqrt{8\ln 2} \approx 2$ instead of 1.4 in equations (21) and (24) as suggested for Gaussian fitting by Condon (1997), neglecting calibration and frame errors, using $\Theta = \Theta_\alpha = \Theta_\delta$ for a circular beam, and noting that the median offset about an input position in 2D is given by the median of a Rayleigh (1880) distribution, we evaluate the expected median positional offset for IMFIT as

$$\text{pos. offset}^{\text{C97}}_{\text{median}} = \sqrt{\ln 4}\,\frac{\Theta}{2A}\,. \tag{36}$$

This estimate is indicated by the dotted curve in each panel of Fig. 8.

Equation (36) suitably encapsulates the positional uncertainties exhibited by both IMFIT and BLOBCAT for unresolved sources. However, for our heavily resolved source, it systematically underestimates the positional uncertainties exhibited by both the Gaussian fit and flood fill approaches. To avoid complexity we do not attempt to explicitly parameterise the increased positional uncertainty displayed for resolved sources. Instead, we have chosen to simply modify the positional uncertainty equations presented by Condon (1997) to use a factor of 1.4 (instead of ~ 2), as presented in equations (21) and (24). These modified equations lead to a more appropriate estimate for the median positional offset,

$$\text{pos. offset}^{\text{BLOBCAT}}_{\text{median}} = \sqrt{\ln 4}\,\frac{\Theta}{1.4\,A}\,, \tag{37}$$

as indicated by the dot-dashed curve in each panel of Fig. 8. The factor of 1.4 was selected empirically to ensure that for Gaussian sources with sizes ranging from unresolved to the heavily resolved source tested, positional uncertainties may be systematically estimated to within $\sim 5\%$ of a beam FWHM. We note that the factor of 1.4 is also suitable for use with IMFIT (see left panels in Fig. 8).

3.2 Linear Polarization

3.2.1 Simulation Setup

We tested BLOBCAT in linear polarization, $L_{\rm RM}$, in a similar manner to that described in § 3.1.1 for total intensity. We tested the same two classes of source, sampling input peak SNRs between $6 - 100\sigma_{\rm RM}$ (cf. equation (5); also § 2.4.2). For comparison, we also tested the performance of IMFIT using both constrained and unconstrained Gaussian fit parameters.

We generated each of the 125 sample images per SNR bin as follows. We assumed an illustrative observational band centred on 1396 MHz with width 200 MHz, split into 25×8 MHz channels. For each frequency channel we obtained two independent noise thumbnails from the master noise image (cf. § 3.1.1), which we used to represent Stokes Q and U noise. A point (or resolved) source with a RM of -100 rad m^{-2}, unresolved in Faraday space, was then suitably injected into each of the Stokes Q and U images across the band. We define the peak SNR of these injected sources as the ratio between their true input peak polarized SB and $\sigma_{\rm RM}$. Using RM synthesis (Brentjens & de Bruyn 2005) and RMCLEAN (Heald et al. 2009), images of $L_{\rm RM}$ were then produced in accordance with equation (1). For each sample we then recovered the peak and integrated SB using both BLOBCAT and IMFIT. We describe the results of these Monte Carlo simulations for SB measurements in § 3.2.2 and for positions in § 3.2.3.

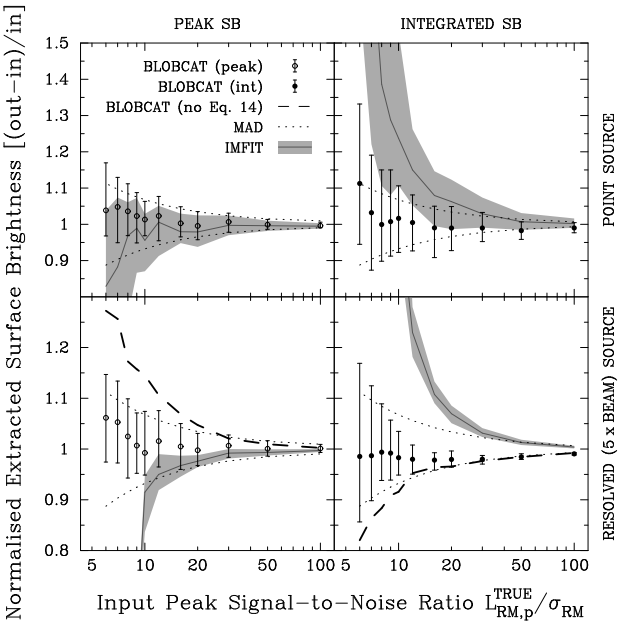

Figure 9. Surface brightness measurement performance of BLOBCAT in linear polarization, $L_{\rm RM}$; the display layout is duplicated from Fig. 6. Fit parameters for IMFIT are unconstrained. No corrections for polarization bias have been applied.

3.2.2 Results and Discussion: Surface Brightness Measurements

We performed our linear polarization Monte Carlo simulations using a range of T_d, T_f, and λ parameter values, finding that the optimal total intensity value of $\lambda = 3.5$ was suitable for use in polarization as well. This behaviour of λ can be understood by comparing profiles through sources embedded within images of total intensity and $L_{\rm RM}$, as presented by Hales et al. (2012). They show that above $T_f = 4$, Gaussian sources embedded within these two environments are very similar in morphology, modulo statistical fluctuations. For this reason, the relevant cross-sectional area for the peak bias correction, H_λ in equation (13), may be obtained for images of $L_{\rm RM}$ using the same value of λ as was recommended for total intensity. Using this value, we found that integrated SB recovery was optimised when flooding down to $T_f = 4.0$, as discussed earlier in § 2.4.2.

In Fig. 9 we present the results of our simulations, where we have executed IMFIT using unconstrained Gaussian fit parameters with a $4\sigma_{\rm RM}$ cut-off fitting threshold (same as T_f). The results obtained from the same simulations by executing IMFIT with constrained point source fits are presented in Fig. 10.

The strong systematic biases exhibited by IMFIT in Fig. 9 suggest that its unconstrained fits are unsuited to the statistical environment of $L_{\rm RM}$. We attribute this to a breakdown in the assumption that sources are superposed

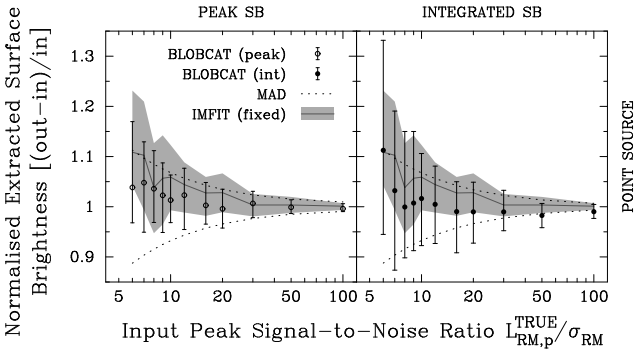

Figure 10. Reproduction of top row of Fig. 9, but here displaying `IMFIT` results for point source fits with angular dimensions fixed to the image resolution.

with Gaussian noise fluctuations, as required to perform robust least squares minimisation. When `IMFIT`'s angular size parameters are fixed to the image resolution, the systematic biases in measured SB for input point sources are diminished, as shown in Fig. 10. Through further experimentation, we found that systematic `IMFIT` biases were unavoidable for all but the most manual, uniquely-constrained fits. Reduction or removal of the $4\sigma_{\rm RM}$ cut-off threshold, used to prevent faint pixels from entering the Gaussian fitting process, was found to worsen systematic trends. We found similar biases to those described above when using `IMFIT` in images of standard linear polarization, L.

In contrast, the results from Fig. 9 indicate that `BLOBCAT`'s measurements of peak and integrated SB are, in effect, systematically unbiased. We justify this claim as follows, beginning with peak SB performance.

The small systematic positive bias exhibited by the recovered peak SB is due to the positive semi-definite nature of $L_{\rm RM} \geqslant 0$; this effect, which is extrinsic to `BLOBCAT`, is known as polarization bias. Because of the intimate relationship that exists between polarization bias and the specifics of observational setup, as elucidated shortly, `BLOBCAT` makes no attempt to correct for this bias. To illustrate the variety and complexity of schemes that may be applicable to different data, we note that corrections designed for L (see Leahy & Fernini 1989) are not suitable for $L_{\rm RM}$ because they are governed by different statistical distributions (Hales et al. 2012). Furthermore, no fixed (unparameterised) correction scheme[7] is suitable for $L_{\rm RM}$, because the statistical properties of $L_{\rm RM}$ are dependent on the underlying observational characteristics of the data such as frequency coverage and channel width (Hales et al. 2012). Instead, more computationally expensive schemes to correct for polarization bias, and potentially Eddington bias (which affects the measured SB of unresolved sources; Eddington 1913), may be required (Hales et al., in preparation). To alleviate polarization bias in `BLOBCAT`'s measurements of peak SB, users must independently apply their own suitably selected correction scheme.

[7] We note that George et al. (2011) recently proposed a fixed correction scheme for $L_{\rm RM}$. As their scheme implicitly assumes a specific observational setup, its applicable parameter space is limited.

`BLOBCAT` appears to accurately recover measurements of integrated SB for unresolved sources, apart from a positive bias exhibited at low input SNR. This latter behaviour is due to polarization bias, which affects sources whose pixel magnitudes are predominantly at low SNR. However, this bias is not of significant consequence because, on average for these sources, their ratios of integrated to peak SB will not deviate significantly from 1. In these cases, their peak values will best represent their flux densities (cf. § 3.1.2; also § 4.1), which need only be corrected for polarization bias in order to deliver systematically unbiased measurements.

Turning to `BLOBCAT`'s measurements of integrated SB for highly resolved sources, their unbiased nature appears to be due to the fortuitous cancellation of two systematic effects. The first of these is the negative bias for resolved sources, as seen earlier for total intensity (lower-right panel of Fig. 6). The second is the positive polarization bias discussed above. We conjecture that the cancellation of these two effects is robust, regardless of the observational setup dictating the specific statistical description displayed by the input $L_{\rm RM}$ (or L) image. Our justification for this assertion is that the dominant statistical differences between images of $L_{\rm RM}$ for different observational setups, or between images of $L_{\rm RM}$ and L, occur below a threshold of $4\sigma_{\rm RM}$ (Hales et al. 2012). Given that `BLOBCAT` ignores data below this cut-off threshold (for our recommended $T_f = 4.0$), we are confident that any systematic blob-extraction differences between these images will be below the noise level.

Regarding SB measurement uncertainties, we mirror the earlier discussion of total intensity uncertainties from § 3.1.2. We note that equations (30) and (32) suitably reflect `BLOBCAT`'s measurement errors in linear polarization, as exhibited by the dotted lines in Fig. 9. We therefore leave these equations unchanged for use in linear polarization analysis.

3.2.3 Results and Discussion: Position Measurements

In Fig. 11 we compare the accuracy of position measurements using both `BLOBCAT` and `IMFIT` in recovering the true input positions for our simulated unresolved and resolved sources. As with SB measurements (§ 3.2.2), we find that unconstrained Gaussian fitting is not appropriate for determining source positions in linear polarization. Following from the discussion for positional measurements in total intensity (§ 3.1.3), we note that `BLOBCAT`'s weighted centroid positions are also suitable for use in linear polarization, as are the uncertainty estimates using equations (21) and (24).

3.3 Complex Blobs

In this section we qualitatively discuss `BLOBCAT`'s performance when analysing blobs that exhibit complex (resolved, non-Gaussian) morphology. We do not seek to quantitatively address this performance due to the lack of clear standardised test sources. Possible examples of complex blobs include supernova remnant shells, extended lobes of radio galaxies, radio relics, and extended Galactic emission; we discuss how these blobs may be automatically identified and flagged for follow-up using `BLOBCAT` in § 4. Other examples include blended blobs that consist of multiple overlapped individual Gaussians; we discuss these in § 4.2.

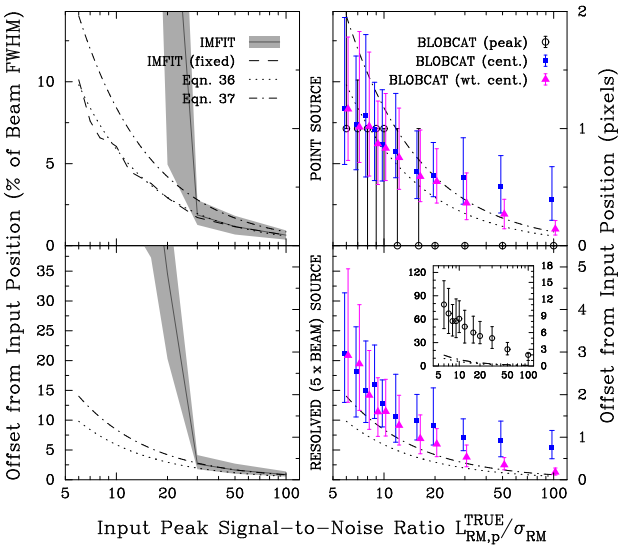

Figure 11. Positional accuracy of `BLOBCAT` and `IMFIT` in linear polarization, $L_{\rm RM}$; the display layout is duplicated from Fig. 8.

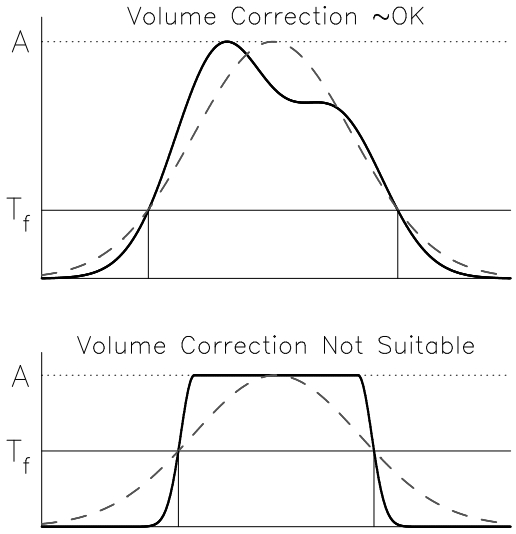

Figure 12. When confronted with a non-Gaussian blob (two arbitrary resolved samples illustrated; solid curves), `BLOBCAT` assumes an idealised Gaussian morphology (dashed curves at equal peak SNR, A) so as to infer the fractional volume remaining unflooded below the cutoff threshold (T_f). If this assumption is particularly poor, as suggested by the example in the lower panel, then the resulting measurement of volume-corrected integrated SB (using equation (17)) may become systematically biased away from the blob's true flux density. For such blobs, the uncorrected measurement of integrated SB is likely to act as a less-biased estimator of true flux density.

For each detected blob, `BLOBCAT` assumes 2D elliptical Gaussian morphology (§ 2.3) so as to infer a debiased peak SB and a corrected integrated SB (§ 2.4). If a detected blob is not of true Gaussian morphology, then its debiased peak SB is unlikely to be significantly affected. This is because use of $\lambda = 3.5$ in calculating the relevant cross-sectional area susceptible to peak bias (using equation (14)) is still likely to be a suitable choice for non-Gaussian blobs. It is more difficult to generalise the systematic manner in which measurements of corrected integrated SB may differ from their true values. The simplest observation is that low SNR blobs are more vulnerable than high SNR blobs to systematic error in their measurements of corrected integrated SB (cf. equation (17)). However, the fraction of blob volume remaining unflooded below T_f will be small for a low SNR blob that is highly-resolved, suggesting that in general, uncorrected integrated SB measurements will be more accurate than corrected integrated SB measurements in estimating flux densities for a majority of complex blobs. We have verified the general statements above by testing `BLOBCAT`'s performance in handling sources with a range of complex morphologies. We find that `BLOBCAT`'s performance for slightly extended non-Gaussian blobs that consist of blended Gaussian components, where the approximation of 2D elliptical Gaussian morphology is poor, is in general poorer than the simulation results presented earlier for pathologically resolved Gaussian blobs. However, alternatives for handling such blobs more suitably in post-processing are available, as discussed in § 4.2. For highly extended non-Gaussian blobs, `BLOBCAT`'s measurements of uncorrected integrated SB are in general quite accurate because the fraction of unflooded blob volume is always very small.

In Fig. 12 we present two sample non-Gaussian blobs in an attempt to illustrate their potential for integrated SB error. Users should judge for themselves whether corrected (S_{int}) or uncorrected ($S_{int}^{\rm OBS}$) measurements of integrated SB best describe the flux densities of their complex blobs; to assist with this decision, `BLOBCAT` reports both values in its output catalogue. If the two values differ by more than a few percent, then the corrected values may be unsuitable, and manual inspection is recommended.

Similarly, users should determine which `BLOBCAT` position measurement is most appropriate for each of their complex blobs; the SNR-weighted centroid may be inappropriate for some blobs. For example, the weighted centroid position for an arc-shaped radio relic (i.e. a crescent moon shape) may be situated beyond the boundaries of its flooded pixels; the raw peak pixel or area (unweighted) centroid position may be more appropriate. To aid users, `BLOBCAT` catalogues all three position measurements. In addition, flags are produced (see § 2.6) so as to indicate whether the centroid positions are situated within or exterior to the flooded pixel confines of each blob.

4 POST-PROCESSING

`BLOBCAT` is designed to produce an output catalogue that details basic properties of blobs in an image. Depending on the nature of the data and the requirements of the user, additional processing may be required to make full use of the catalogue.

In this section we highlight two such examples of post-processing. We first consider a selection procedure for determining which SB measurement (peak or integrated) best describes the flux density of a blob. We then consider a procedure for identifying and analysing blobs that exhibit non-Gaussian morphologies.

4.1 Blob Flux Densities

The choice of whether to represent a blob's flux density by its measured peak or integrated SB is equivalent to asking whether the blob is unresolved or not. If it is unresolved then the peak SB should be used (explained as follows; note also Appendix A), while for resolved blobs it is the integrated SB that should be used.

The user is responsible for selecting which of the measurements of peak or integrated SB best represent the true flux density for each detected blob. We do not automate this process for the same reason that Gaussian fitting tasks such as `IMFIT` do not, namely that noise features adjacent to faint, unresolved sources may render integrated SB measurements less likely to represent true flux densities than peak SB measurements.

If a user is only interested in a small number of blobs, then as with `IMFIT`, more attention can be paid to each individual fit so as to minimise potential fitting errors, for example through fitting constraints in `IMFIT` or perhaps suitable pixel masking prior to running `BLOBCAT`. For such carefully fitted blobs, their integrated SB measurements may be used to represent their true flux densities, even if they are faint or unresolved. However, for large sample sizes (e.g. for a survey), it is impractical to consider implementation of such manual, or perhaps even machine-learning enabled, fitting procedures. Indeed, attempting to manually fit each source in a survey may inadvertently bias the resulting flux density measurements due to subjectivity on behalf of the user.

Instead, a more appropriate strategy may be initiated by taking the ratio between integrated to peak SB measurements for each blob, so as to characterise the global variance in this ratio as a function of measured SNR. By considering the parameter space populated by noise-affected blobs with $S_{int} < S_p$, an envelope can be designed as a function of SNR to select which of the blobs with $S_{int} > S_p$ are likely to be similarly affected by noise. Only those blobs with ratios in excess of the envelope criterion may be deemed resolved, and in turn have their flux densities represented by their integrated SB measurements. All other blobs should have their flux densities represented by their peak SB measurements. This strategy has been employed for `IMFIT`-based surveys of total intensity, e.g. Huynh et al. (2005); application to total intensity and linear polarization surveys with `BLOBCAT` will be presented by Hales et al. (in preparation).

If a blob is resolved, then an estimate of its deconvolved size may be obtained directly from its integrated to peak SB ratio (via division of equation (8) by equation (4)), namely

$$\frac{S_{int}}{S_p} = \frac{\psi_r \, \psi_s}{\Theta_{\rm maj} \, \Theta_{\rm min}} \ , \tag{38}$$

where the deconvolved angular size can be estimated using the geometric mean as $\psi_{\rm deconv} \approx \sqrt{\psi_r \, \psi_s - \Theta_{\rm maj} \, \Theta_{\rm min}}$. Again, illustrations of this procedure are available in total intensity using `IMFIT` (Huynh et al. 2005), and will be presented for total intensity and linear polarization with `BLOBCAT` by Hales et al. (in preparation).

4.2 Blob Decomposition

`BLOBCAT` assumes that isolated blobs are of Gaussian morphology in order to catalogue their properties. This assumption will work well for images that are sparsely populated (i.e. not confusion limited) with Gaussian sources. However, if complex blobs are present (cf. § 3.3) this assumption may not always be suitable, requiring additional processing of the complex objects so as to suitably characterise their properties. Before commenting on this processing, we briefly outline a simple procedure by which complex blobs may be first identified.

In equation (34) we defined the parameter $R^{\rm EST}$, which estimates the size of a detected blob in units of the sky area covered by an unresolved Gaussian blob with the same peak SB. If $R^{\rm EST}$ is large, it indicates that a blob is unlikely to be unresolved.

To illustrate how this parameter may be used to identify potentially complex blobs for follow-up, we preview the general processing steps performed by Hales et al. (in preparation) to catalogue sources in radio-wavelength images of total intensity and linear polarization; details of these images are not pertinent to the discussion here, apart from noting that they consist mostly of compact sources (i.e. there are no wide-spread extended image features). Hales et al. (in preparation) find that a value of $R^{\rm EST} > 1.4$ is well-suited for automatically flagging complex blobs. Gaussian fits are attempted for each of these flagged complex blobs with `IMFIT` to determine which ones are likely to consist of single or multiple overlapped (blended) Gaussians. This procedure is semi-automated to require only two initial manual inputs to `IMFIT`: the number of potentially overlapped Gaussians, and their cursory positions. We note here that standard digital imaging techniques such as the Laplacian of Gaussian operation (e.g. Sonka, Hlavac & Boyle 2008) which is implemented within the `AEGEAN` algorithm (Hancock et al. 2012), blob decomposition algorithms such as `CLUMPFIND` (Williams et al. 1994), or the widely used Watershed transform (Roerdink & Meijster 2000), may be well suited to performing this step automatically. Hales et al. (in preparation) preserve the original `BLOBCAT` measurements for those blobs that are best fit by a single Gaussian. For each blob identified as being blended, they replace its original `BLOBCAT` catalogue entry with multiple `IMFIT` entries for each individual Gaussian component identified. Remaining from this procedure are a small number of extended, non-Gausssian blobs that cannot be adequately refit using `IMFIT` (as identified due to their large fitting residuals; we note here that image artefacts may also be included in this list, though too many artefacts could indicate undervaluation of rms noise estimates). For each of these remaining blobs, Hales et al. (in preparation) preserve the original `BLOBCAT` measurements and perform a final manual inspection to determine which of the integrated SB measurements should be used to represent the blob's flux density (uncorrected or corrected; § 3.3).

We envisage that the above procedure may be quickly and easily replicated for future surveys. By performing Gaussian fitting for only those blobs that `BLOBCAT` indicates may be complex, it should be possible to robustly and automatically catalogue all but the most non-Gaussian of sources in an image.

5 SUMMARY AND CONCLUSIONS

We have described `BLOBCAT`, an algorithm designed to identify and catalogue blobs in a 2D FITS image of Stokes I

intensity or linear polarization (L or $L_{\rm RM}$). Utilising a Gaussian morphology assumption and two key bias corrections, `BLOBCAT` equips its core flood fill algorithm with the tools necessary to perform robust SB measurement.

Written in `Python`, `BLOBCAT` is easy to use and easy to modify. It is well-suited to analysis of large blind surveys, requiring little manual intervention for images sparsely populated with unresolved and resolved Gaussian sources, and having the ability to account for spatial variations in both image sensitivity and bandwidth smearing. To indicate `BLOBCAT`'s ability to swiftly analyse data, we note that Hales et al. (in preparation) produce a catalogue of ~ 1000 blobs from an image with $\sim 10\,000 \times 10\,000$ pixels, including the use of equal-sized rms and bandwidth smearing images, in less than 60 seconds on a standard desktop computer. While source extractors built around Gaussian fitting routines are competitive with `BLOBCAT` in this raw computing time, though such comparison is implementation-dependent, subsequent overheads associated with manual source inspection may be greatly minimised when using the latter. This is because unresolved and resolved Gaussian blobs are automatically and accurately processed by `BLOBCAT`, requiring only non-Gaussian blobs to be manually addressed.

Accurate estimates of background rms noise are required to ensure robust and accurate operation of `BLOBCAT`. We described a simple, objective, and automated procedure by which these estimates may be obtained, which makes use of local background mesh calculations. We note that this procedure may be used to estimate background rms noise for use with any source extractor, not just `BLOBCAT`.

We have demonstrated the performance of `BLOBCAT` through Monte Carlo simulations of unresolved and resolved Gaussian sources. We benchmarked this performance against that of standard Gaussian fitting, finding comparable results in total intensity, and vastly superior results in linear polarization. Our simulations indicate that Gaussian fitting is inappropriate for use in linear polarization for all but the most manually-constrained of fits. `BLOBCAT` contains at present the only algorithm capable of robustly cataloguing accurate flux densities for resolved or extended sources in linear polarization, without incurring significant systematic biases.

In closing, we note that `BLOBCAT` may be suitable for cautious application to image data at non-radio wavelengths, such as optical, provided that the flooding SNR cutoff is set to a value high enough to avoid measurement systematics induced by low-SNR statistics. Optical pixel shot noise (the Poisson regime) is non-Gaussian at low-SNR and limits to Gaussianity at higher SNR, much like the statistics of linear polarization that can be accommodated by `BLOBCAT`. Modification of `BLOBCAT`'s algorithms may be required to account for wavelength- and instrument-specific descriptions of point spread functions and pixellation errors.

The `BLOBCAT` program, supplemented with test data to illustrate its use, is available electronically through the World Wide Web at: `http://blobcat.sourceforge.net/` .

ACKNOWLEDGMENTS

We thank the following for helpful discussions and feedback: Tim Cornwell, Mark Calabretta, Andrew Hopkins, Elizabeth Mahony, Paul Hancock, Greg Madsen, and Jay Banyer. We thank the anonymous referee for helpful comments that led to the improvement of this paper. C. A. H. acknowledges the support of an Australian Postgraduate Award and a CSIRO OCE Scholarship. The Centre for All-sky Astrophysics is an Australian Research Council Centre of Excellence, funded by grant CE11E0090.

REFERENCES

Bertin E., Arnouts S., 1996, A&AS, 117, 393
Bondi M., et al., 2003, A&A, 403, 857
Brentjens M. A., de Bruyn A. G., 2005, A&A, 441, 1217
Bridle A. H., Greisen E. W., 1994, AIPS Memo No. 87, NRAO, Charlottesville, VA
Bridle A. H., Schwab F. R., 1999, Synthesis Imaging in Radio Astronomy II, 180, 371
Briggs D. S., Cornwell T. J., 1992, ASPC, 25, 170
Briggs D. S., 1995, PhD thesis, The New Mexico Institute of Mining and Technology
Calabretta M. R., Greisen E. W., 2002, A&A, 395, 1077
Condon J. J., 1997, PASP, 109, 166
Condon J. J., Cotton W. D., Greisen E. W., Yin Q. F., Perley R. A., Taylor G. B., Broderick J. J., 1998, AJ, 115, 1693
Cornwell T., Fomalont E. B., 1999, Synthesis Imaging in Radio Astronomy II, 180, 187
David H. A., Nagaraja H. N., 2003, Order Statistics (3rd ed.). John Wiley & Sons, NJ
Deboer D. R., et al., 2009, IEEE Proceedings, 97, 1507
Eddington A. S., 1913, MNRAS, 73, 359
Ellingson S. W., Clarke T. E., Cohen A., Craig J., Kassim N. E., Pihlstrom Y., Rickard L. J., Taylor, G. B., 2009, IEEE Proceedings, 97, 1421
Fishkin K. P., Barsky B. A., 1985, in Magnenat-Thalmann N., Thalmann D., ed., Proc. Graphics Interface '85. Springer-Verlag, Tokyo, p. 56
Fruchter A. S., Hook R. N., 2002, PASP, 114, 144
Gaensler B. M., Landecker T. L., Taylor A. R., POSSUM Collaboration, 2010, BAAS, 42, 515
George S. J., Stil J. M., Keller B. W., 2011, PASA, in press (arXiv:1106.5362)
Gooch R., 1996, Astronomical Data Analysis Software and Systems V, 101, 80
Grant J. K., Taylor A. R., Stil J. M., Landecker T. L., Kothes R., Ransom R. R., Scott D., 2010, ApJ, 714, 1689
Greisen E. W., 1983, AIPS Memo No. 27, NRAO, Charlottesville, VA
Hales C. A., Gaensler B. M., Chatterjee S., van der Swaluw E., Camilo F., 2009, ApJ, 706, 1316
Hales C. A., Gaensler B. M., Norris R. P., Middelberg E., 2012, MNRAS, in press
Hancock P. J., Murphy T., Gaensler B. M., Hopkins A., Curran J. R., 2012, MNRAS, 422, 1812
Heald G., Braun R., Edmonds R., 2009, A&A, 503, 409
Hills R. E., Kurz R. J., Peck A. B., 2010, Proc. SPIE, 7733, 773317
Holwerda B. W., 2005, arXiv:astro-ph/0512139
Hopkins A. M., Afonso J., Chan B., Cram L. E., Georgakakis A., Mobasher B., 2003, AJ, 125, 465

Huynh M. T., Jackson C. A., Norris R. P., Prandoni, I., 2005, AJ, 130, 1373
Ibar E., Ivison R. J., Biggs A. D., Lal D. V., Best P. N., Green D. A., 2009, MNRAS, 397, 281
Johnson N. L., Kotz S., 1970, Distributions in Statistics: Continuous Univariate Distributions−1. Houghton Mifflin, NY
Johnston S., et al., 2008, Experimental Astronomy, 22, 151
Jonas J. L., 2009, IEEE Proceedings, 97, 1522
Joye W. A., Mandel E., 2003, Astronomical Data Analysis Software and Systems XII, 295, 489
Leahy P., Fernini I., 1989, VLA Scientific Memorandum No. 161, NRAO
Lieberman H., 1978, in SIGGRAPH '78 Proceedings. ACM, New York, p. 111
Mauch T., Murphy T., Buttery H. J., et al., 2003, MNRAS, 342, 1117
Murphy T., Mauch T., Green A., Hunstead R. W., Piestrzynska B., Kels A. P., Sztajer P., 2007, MNRAS, 382, 382
Norris R. P., et al., 2006, AJ, 132, 2409
Norris R. P., et al., 2011, PASA, 28, 215
Oosterloo T., Verheijen M. A. W., van Cappellen W., Bakker L., Heald G., Ivashina M., 2009, PoS(SKADS2009), 070
Oppermann N., Robbers G., Ensslin T. A., 2011, arXiv:1107.2384
O'Sullivan S., Stil J., Taylor A. R., Ricci R., Grant J. K., Shorten K., 2008, in POS, Proc. 9th Eur. VLBI Network Symp. for Radio Astron., p. 107
Pach J., Agarwal P. K., 1995, Combinatorial Geometry. John Wiley & Sons, NY
Pence W. D., Chiappetti L., Page C. G., Shaw R. A., Stobie E., 2010, A&A, 524, A42
Perley R. A., Chandler C. J., Butler B. J., Wrobel J. M., 2011, ApJL, 739, L1
Rayleigh J. W. S., 1880, Philosophical Magazine, 5th Series, 10, 73
Roerdink J. B. T. M., Meijster A., 2000, Fundamenta Informaticae, 41, 187
Rottgering H., et al., 2010, PoS(ISKAF2010), 050
Rudnick L., 2002, PASP, 114, 427
Rudnick L., Brown S., 2009, AJ, 137, 145
Sault R. J., Teuben P. J., Wright M. C. H., 1995, Astronomical Data Analysis Software and Systems IV, 77, 433
Shi H., Liang H., Han J. L., Hunstead R. W., 2010, MNRAS, 409, 821
Sonka M., Hlavac V., Boyle R., 2008, Image Processing, Analysis, and Machine Vision (3rd ed.). Thompson Learning, Toronto
Subrahmanyan R., Ekers R. D., Saripalli L., Sadler E. M., 2010, MNRAS, 402, 2792
Taylor A. R., Salter C. J., 2010, in Kothes R., Landecker T. L., Willis A. G., eds, ASP Conf. Ser. Vol. 438, The Dynamic Interstellar Medium: A Celebration of the Canadian Galactic Plane Survey. Astron. Soc. Pac., San Francisco, p. 402
Taylor A. R., et al., 2007, ApJ, 666, 201
Tukey J. W., 1977, Exploratory Data Analysis. Addison-Wesley, Reading, MA
Vaillancourt J. E., 2006, PASP, 118, 1340
Williams J. P., de Geus E. J., Blitz L., 1994, ApJ, 428, 693; see also http://www.ifa.hawaii.edu/users/jpw/clumpfind.shtml
Wootten A., Thompson A. R., 2009, IEEE Proceedings, 97, 1463

APPENDIX A: PIXELLATION ERROR

In radio synthesis imaging, the number of pixels per resolution element (synthesised beam) can be adjusted after the original observations have been made. This is because raw data are obtained in the Fourier plane, enabling post facto over-sampling of data in the image plane. By comparison, optical observations are often under-sampled in the image plane, requiring ingenious methods to utilise their full resolution (e.g. the Drizzle algorithm by Fruchter & Hook 2002).

In this Appendix we present implications for SB measurements when sampling a radio image with insufficient pixels. We use the term 'pixellation error' to refer specifically to the systematic undervaluation of peak SB measurements due to imaging and fitting effects. We focus on the pixellation error exhibited by two methods of peak SB measurement for unresolved sources. We first derive a relationship for the pixellation error exhibited by measurements of observed (raw) peak SB. We then compare this peak pixel error to that exhibited by the fitted peak of a 2D parabola, where the fit is obtained using a 3×3 pixel array about the raw peak pixel (e.g. as implemented in the `MIRIAD` task `MAXFIT`). We conclude by commenting on the manner in which image pixellation affects measurements of integrated SB.

In conventional radio synthesis imaging, the sky is assumed to be represented by delta functions; each image pixel is thus a spot sample, as opposed to other sky representations such as piecewise-constant pixels, which require integrals over regions to be computed. To represent the visibility data, sources in images deconvolved using the iterative CLEAN algorithm will be of the form (Briggs & Cornwell 1992; Briggs 1995)

$$S^{\rm OBS}(x,y) = [BF * SRC * BEAM](x,y)\,, \qquad \text{(A1)}$$

where $S^{\rm OBS}(x,y)$ is the observed source SB distribution at pixel coordinate (x,y), the asterisks indicate convolution, BF is a basis function that depends on whether the source is centred directly on a pixel or not, SRC represents the clean component model of the source, and $BEAM$ is the restoring beam. We assume that $BEAM$ is Gaussian.

We define $\varepsilon^{\rm OBS} = S_p^{\rm OBS}/S_p^{\rm TRUE}$ as the fraction of true peak SB observed within the peak pixel of an unresolved source. We assume N_α and N_δ pixels per projected resolution element such that a pixel dimension is $\Theta_\alpha/N_\alpha \times \Theta_\delta/N_\delta$; here, Θ_α and Θ_δ are the major and minor FWHMs that characterise the image resolution (see introductory remark in § 2.3), as projected along the RA and Dec axes of an image (see equations (20) and (25)).

When the true peak for an unresolved 2D elliptical Gaussian is centred directly on a pixel, which we denote the 'on-pixel' case, both the BF and SRC terms in equation (A1) are given by delta functions. The source SB distribution is therefore described by an unattenuated 2D elliptical Gaussian with $\varepsilon^{\rm OBS}_{\rm on\text{-}pixel} = 1$, regardless of the values of N_α and N_δ.

When the true peak is centred half-way between pixel

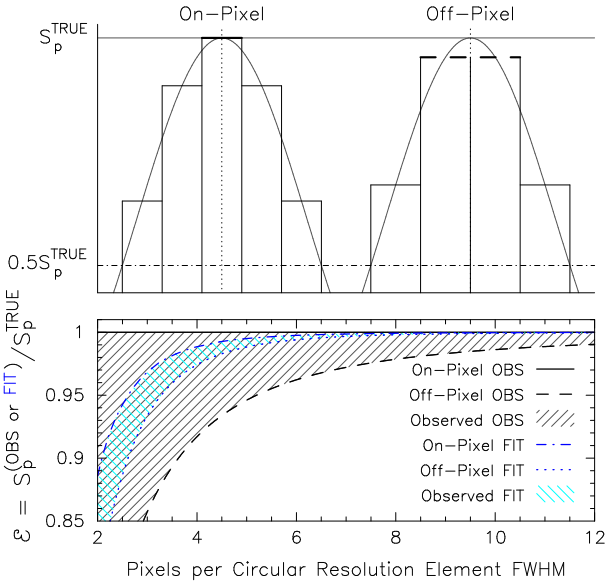

Figure A1. Peak surface brightness underestimation due to pixellation; we term this pixellation error. Shown in the upper panel are two unresolved 1D Gaussians with true peak brightness $S_p^{\rm TRUE}$, sampled with 5 (left) and 4 (right) pixels per FWHM. Their true peaks are centred directly on (left) or half-way between (right) pixels. The observed central peak pixel(s) underestimates the true peak brightness of an unresolved 2D elliptical Gaussian by $\varepsilon^{\rm OBS}$, as illustrated in the lower panel for the best-case (true peak centred on a 2D pixel; solid curve), worst-case (true peak placed at the intersection of 4 pixels; dashed curve), and intermediate-case (right-slant shaded) pixellation of a circular resolution element (i.e. assuming $N_\alpha = N_\delta$). Similarly, the underestimate exhibited by the fitted peak of a 2D parabola, $\varepsilon^{\rm FIT}$, is illustrated in the lower panel for the best-case (dot-dashed curve), worst-case (dotted curve), and intermediate-case (left-slant shaded) pixellation scenarios.

centres, which we denote the 'off-pixel' case, SRC is again a delta function (representing a point source) and $BEAM$ is a Gaussian, but now BF must consist of a sinc function in order to represent the visibility data for a shifted delta function. We find that $\varepsilon^{\rm OBS}_{\rm off\text{-}centre}$ is therefore given by

$$\varepsilon^{\rm OBS}_{\rm off\text{-}centre} = \frac{1}{S_p^{\rm TRUE}} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{\sin(\pi l)}{\pi l}\frac{\sin(\pi m)}{\pi m} \times S_p^{\rm TRUE} \exp\left\{ -4\ln[2] \left[\frac{(x_{1/2}-l)^2}{N_\alpha^2} + \frac{(y_{1/2}-m)^2}{N_\delta^2} \right] \right\} dl\, dm\,, \quad \text{(A2)}$$

evaluated at $x_{1/2} = y_{1/2} = 0.5$.

In Fig. A1 we display $\varepsilon^{\rm OBS}$ for the on- and off-pixel cases from above; to conform with visual expectations, in the upper panel we plot 1D source profiles and their corresponding 1D pixel values by using a simplified 1D version of equation (A2) (for which only one integral is required). When the underlying true peak for an unresolved source is centred (in 2D) part-way between the on- and off-pixel cases, $\varepsilon^{\rm OBS}$ is given by a value between these two solutions, as illustrated by the shading in the lower panel of Fig. A1. We note that the effect of the sinc function in our off-pixel analysis is essentially negligible, only affecting the plotted curves closer to ~ 1 pixel per FWHM. Nevertheless, we have included the calculation for completeness.

In principle, the pixellation error exhibited by measurements of observed peak SB ($\varepsilon^{\rm OBS}$) may be minimised by imaging with a large number of pixels per resolution element. However, in practice, limited computing resources will often prevent the production or subsequent analysis of such heavily sampled images. Rather than increasing the image sampling N_α and N_δ, the accuracy of peak SB measurements may be increased by performing a fit to the peak value using a 2D parabola; we denote these fitted peak measurements $S_p^{\rm FIT}$. To demonstrate this increased accuracy, in Fig. A1 we illustrate the pixellation error exhibited by 2D parabolic fitting, which we define as $\varepsilon^{\rm FIT} = S_p^{\rm FIT}/S_p^{\rm TRUE}$. We note that our $\varepsilon^{\rm FIT}_{\rm on\text{-}pixel}$ and $\varepsilon^{\rm FIT}_{\rm off\text{-}pixel}$ curves in Fig. A1 were obtained analytically; for brevity we will not reproduce the straightforward derivation of $S_p^{\rm FIT}$ here. This derivation involves evaluating raw pixel intensities from either spot samples from a 2D Gaussian for the on-pixel case, or evaluating equation (A2) at different pixel positions for the off-pixel case, then performing least squares to solve for an overdetermined system of linear equations (6 unknown fit parameters and 9 constraining pixels).

Both $S_p^{\rm OBS}$ and $S_p^{\rm FIT}$ exhibit pixellation error; the latter measure of peak SB is more accurate. To limit pixellation error to within 1% using $S_p^{\rm OBS}$, at least 12 pixels per FWHM are required; for $S_p^{\rm FIT}$, this number falls to around 5. We suggest that observers estimate the degree to which their peak SB measurements may be in error due to pixellation, and incorporate this into their error budgets. In `BLOBCAT`, which catalogues fitted peak SB values ($S_p^{\rm FIT}$), this is implemented using a pixellation error parameter which we define as $\Delta S^{\rm PIX} = (1 - \varepsilon^{\rm FIT}_{\rm off\text{-}centre})$; this parameter is applied in equation (30). We note that inclusion of this parameter will tend to (slightly) over-estimate peak SB errors for resolved sources; we see this as more appropriate than underestimating peak SB errors for point sources because this error is unlikely to be relevant for resolved sources (where the integrated SB represents the flux density; see § 4.1).

Finally, we note that integrated SB measurements are less affected by pixellation error than peak pixels. This is because integrated SB is conserved when summing over multiple pixels. This conservation is limited only by noise fluctuations and the ratio between the peak SNR of a source and the flood fill cutoff. To illustrate this limitation, consider a faint unresolved source situated in a heavily pixellated image (i.e. where N_α and N_δ are small). The profile of this source will be poorly mapped by the pixels, rendering `BLOBCAT`'s integrated SB measurement (via equation (16)) vulnerable to negative bias. However, in general this vulnerability will not be an issue because it is the peak SB that is the important value for unresolved sources (see § 4.1).

APPENDIX B: `BLOBCAT` INPUTS

For completeness, a full list of program input arguments to `BLOBCAT` is presented below. Note that not all arguments may be required for analysis (see § 2.5−2.6; see also the default values provided in the code). For example, if errors are not required (or are not suitably defined for a particu-

lar observational scenario), the input arguments relating to errors below may be ignored. (Conversely, new input arguments may be easily defined by the user and incorporated into `BLOBCAT`.)

Argument 1: `SB_image.fits`
FITS image of surface brightness in Stokes I intensity (or Stokes Q, U, or V intensities under limited conditions) or linear polarization (L or $L_{\rm RM}$); see § 2.1.1.

Argument 2: `rmsval`
Uniform (spatially-invariant) background rms noise level within SB image. This is required if Argument 3 is not provided.

Argument 3: `rmsmap`
FITS image of background rms noise; see § 2.1.2.

Argument 4: `bwsval`
Uniform (spatially-invariant) level of bandwidth smearing present in the SB image. This is required if Argument 5 is not provided. To ignore bandwidth smearing, this value should be set to 1.

Argument 5: `bwsmap`
FITS image of background rms noise; see § 2.1.3.

Arguments 6-8: `bmaj, bmin, bpa`
Image resolution (beam) parameters; these are only required if image header items are incorrect or incomplete (at present, beam parameters are not standard FITS headers).

Arguments 9-10: `dSNR, fSNR`
SNR thresholds for blob detection (T_d) and flooding cutoff (T_f); see § 2.2.

Argument 11: `pmep`
Maximum estimated peak SB attenuation due to pixellation error (see Appendix A); defined here as the maximum anticipated value of $(1 - \varepsilon^{\rm FIT}_{\rm off\text{-}centre})$. When set to a value greater than 0, this parameter will ensure that sources with raw observed peak SB less than the nominated detection threshold ($S_p^{\rm OBS} < T_d$), yet fitted peak SB greater than this threshold ($S_p^{\rm FIT} \geqslant T_d$), will be accepted into the catalogue. If ignored, `pmep` will default to 1, causing `BLOBCAT` to check all blobs with $S_p^{\rm OBS} \geqslant T_f$ for catalogue acceptance (though this will increase `BLOBCAT`'s run-time, particularly if T_d and T_f differ greatly in magnitude).

Arguments 12-13: `cpeRA, cpeDec`
Phase calibrator positional error in RA ($\sigma_{\alpha,\rm cal}$) and Dec ($\sigma_{\delta,\rm cal}$); see § 2.6.

Argument 14: `SEM`
Standard error of the mean of the variation in the phase corrections resulting from phase self-calibration ($\sigma_{\rm SEM}$), which is used to calculate $\sigma_{\rm frame}$; see § 2.6.

Argument 15: `pasbe`
Percentage absolute SB error resulting from calibration ($\Delta S^{\rm ABS}$); see § 2.6.

Argument 16: `pppe`
Percentage peak SB pixellation error ($\Delta S^{\rm PIX}$); see § 2.6 and Appendix A.

Argument 17: `cb`
Average clean bias correction ($\Delta S^{\rm CB} \geqslant 0$); see § 2.6.

Argument 18: `lamfac`
λ factor for peak SB bias correction; see § 2.4.1.

Argument 19: `visArea`
Option to calculate visibility areas (can increase program run-time by more than an order of magnitude); see § 2.6.

Arguments 20-22: `minpix, maxpix, pixdim`
Minimum and maximum accepted blob sizes in pixels, and minimum number of pixels in RA/Dec dimensions for accepted blobs (useful for filtering out easily-identified image artefacts).

Argument 23: `edgemin`
Edge buffer in pixels; if flood fill attempts to enter this buffer zone, the blob is rejected (and reported to the user).

Arguments 24-25: `write, hfill`
Options to write flooded blobs to an output FITS file and to set the blob highlight value; see § 2.7.

Arguments 26-27: `kvis, ds9`
Options to write an output `kvis` or `ds9` overlay file; see § 2.7.

Arguments 28-29: `plot, plotRng`
Option to produce a diagnostic screen plot displaying flooded blobs in the SB image, and additional option to specify this plot's dynamic range.

This paper has been typeset from a TEX/ LATEX file prepared by the author.